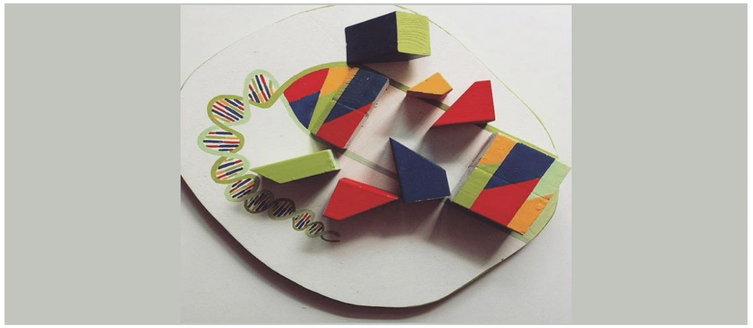
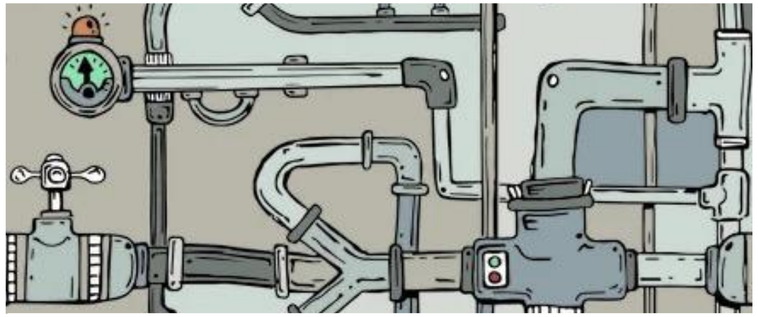
St. Jude Children's Research Hospital | DNB Bioinformatics core





Bulk RNA-Seq Data Standards and Processing Pipeline

Antonia Chroni, PhD
Senior Bioinformatics Research Scientist
DNB Bioinformatics Core

September 19, 2024

Bioinformatics core, Department of Developmental Neurobiology

Providing advanced bioinformatic services for investigators to leverage omics data

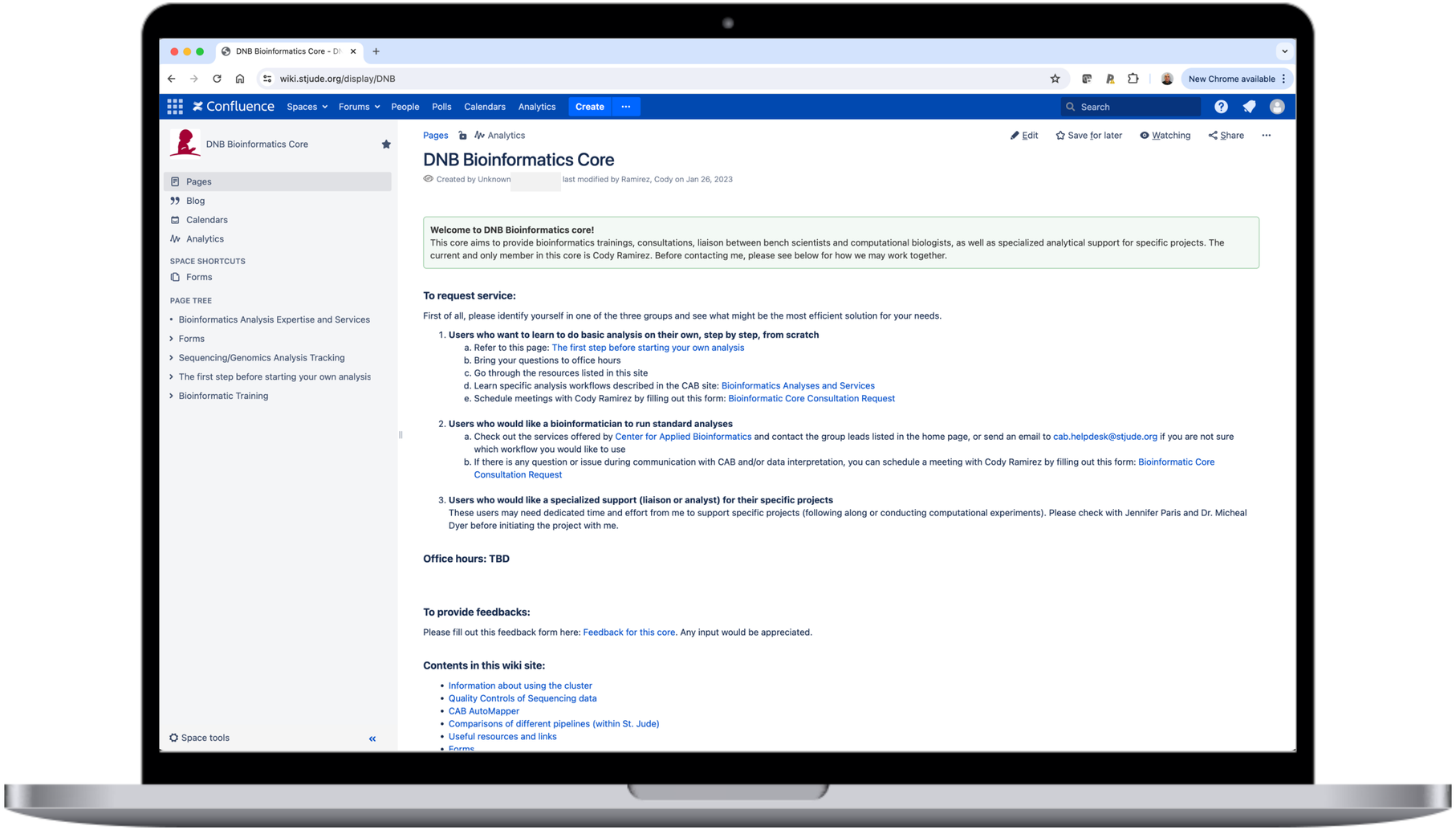
- Project analysis
- Benchmarking

- Training

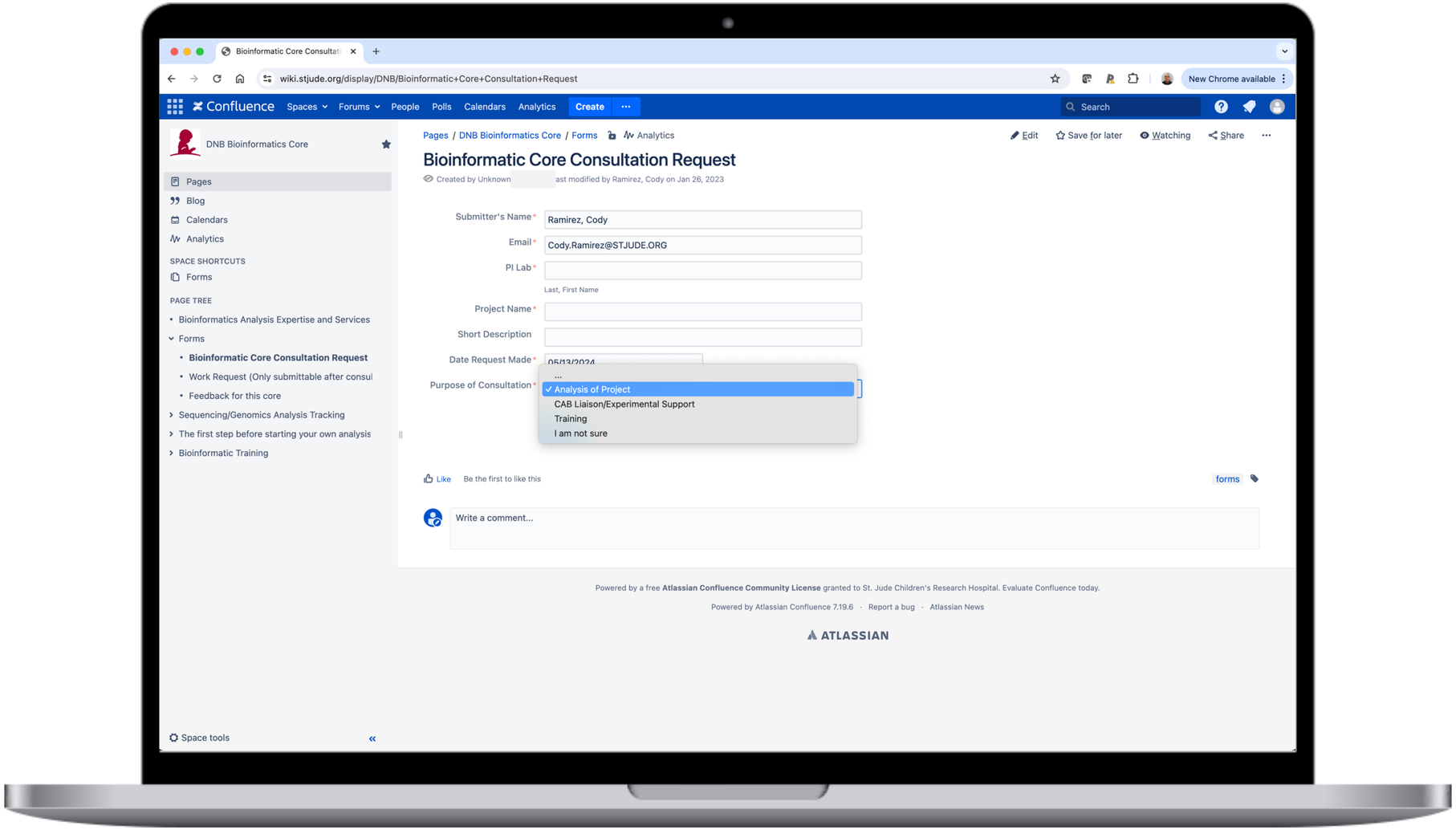
- CAB liaison
- Consultation

- Innovation

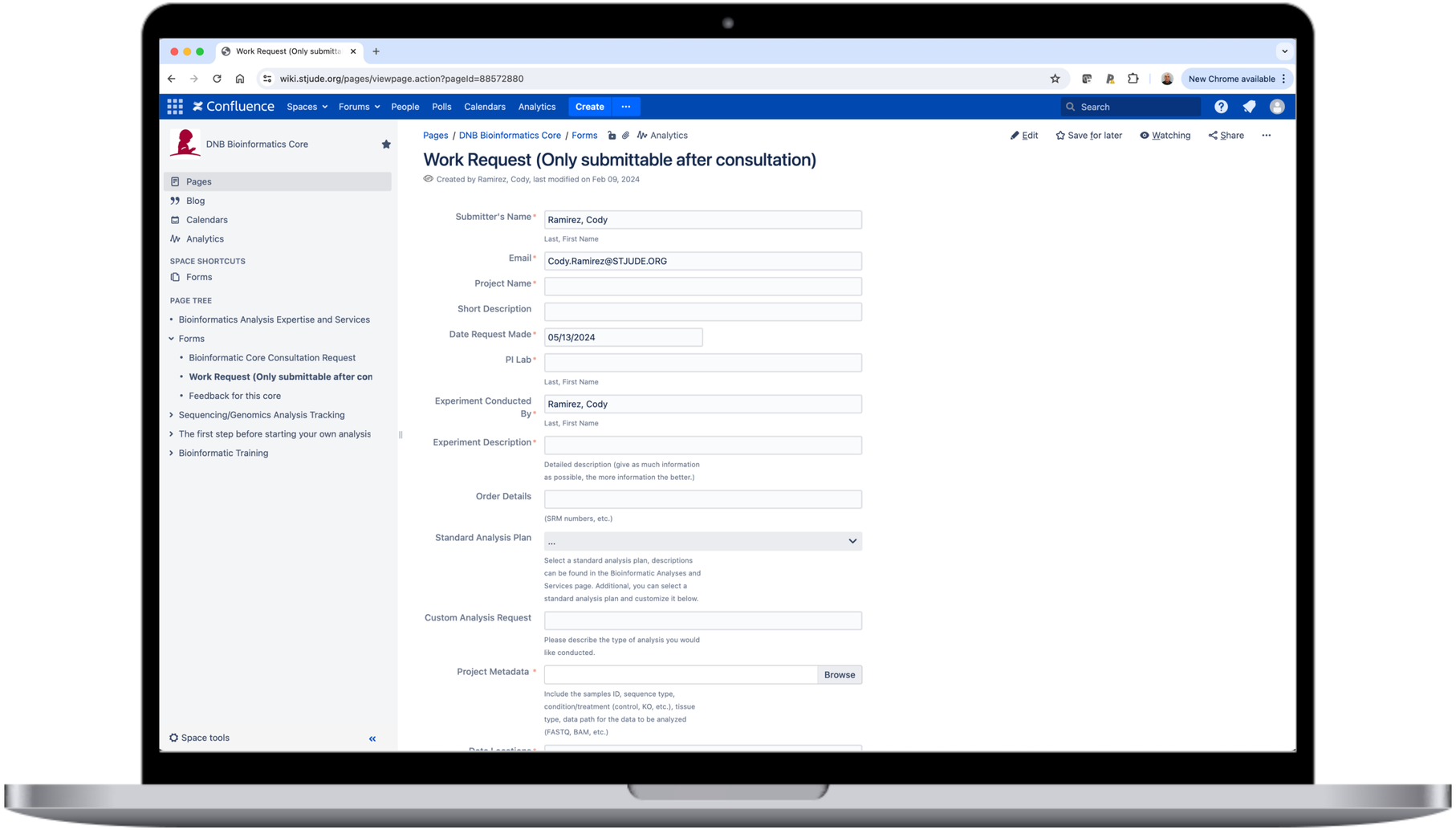
Let us help you! – https://wiki.stjude.org/display/DNB



Talk with us! – Consultation Request



Collaborate with us! – Work Request



Bioinformatics core, Department of Developmental Neurobiology

Providing advanced bioinformatic services for investigators to leverage omics data

- Project analysis
- Benchmarking

- Training

- CAB liaison
- Consultation

- Innovation

GitHub

Bioinformatics core, Department of Developmental Neurobiology

Providing advanced bioinformatic services for investigators to leverage omics data

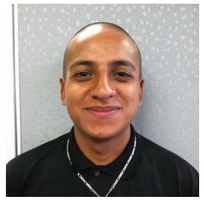
- Project analysis
- Benchmarking

- Training

- CAB liaison
- Consultation

- Innovation

GitHub



Cody Alexander Ramirez, PhD
Senior Bioinformatics Research Scientist
Core Director
Boston, Massachusetts

Antonia Chroni, PhD
Senior Bioinformatics Research Scientist
New York, New York

Sharon Freshour, PhD
Bioinformatics Research Scientist
St. Louis, Missouri

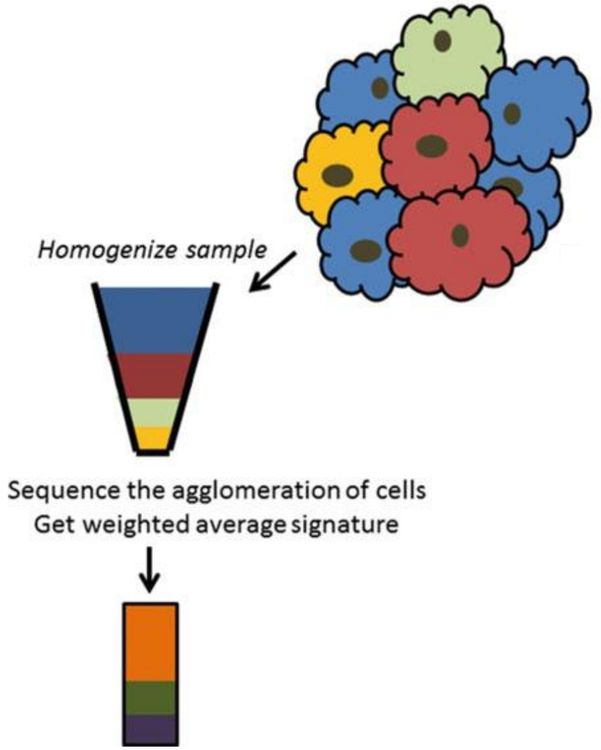
Asha Jacob Jannu, PhD
Bioinformatics Research Scientist
Indianapolis, Indiana

Agenda

1. What is bulk RNA-Sequencing?
2. What can I investigate with RNA-Seq?
3. Key Steps in RNA-Sequencing
   - Library preparation
   - Sequencing technology
   - Data analysis
4. CAB AutoMaper RNA-Seq pipeline and QC metrics Standards
5. Downstream analysis workflow

What is bulk RNA-Sequencing?

- RNA from a mixed population of cells or tissues
- Sequencing to determine the quantity and sequence of RNA transcripts
- **Average gene expression profile for the entire sample,** reflecting the combined expression patterns of all cells within that sample

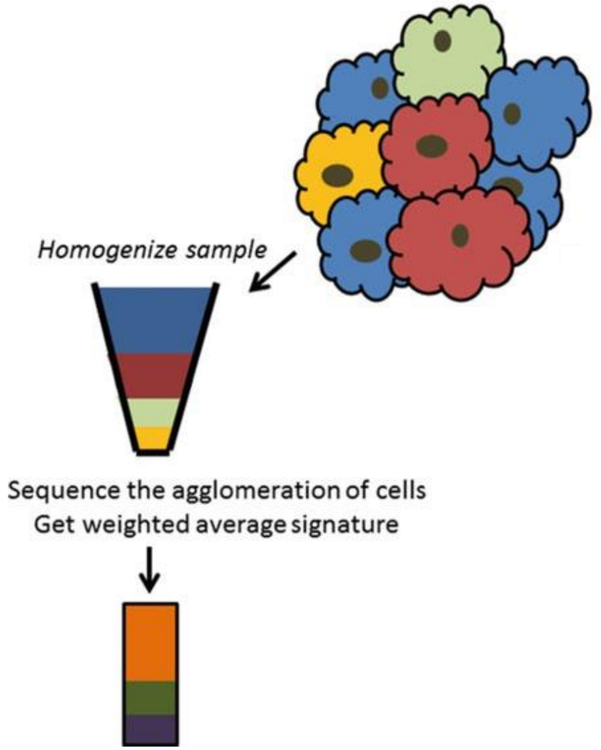


Yu *et al*., 2021. *Translational Bioinformatics for Therapeutic Development*, 143-175.

What can I investigate with RNA-Seq?

Applications

- Differential gene expression
- Isoform switching
- New genes, and transcripts
- New transcriptomes
- Variants
- Allele-specific expression
- Alternative splicing
- Gene fusion
- RNA editing
- Exploration of non-model-organism transcriptomes



Yu *et al*., 2021. *Translational Bioinformatics for Therapeutic Development*, 143-175.

Advantages and Limitations

- **High Throughput**: Provides comprehensive data on gene expression for thousands of genes simultaneously.
- **Quantitative**: Offers precise measurements of gene expression levels.
- **Discovery-Oriented**: Useful for discovering new genes, isoforms, and regulatory elements.

Advantages and Limitations

- **High Throughput**: Provides comprehensive data on gene expression for thousands of genes simultaneously.
- **Quantitative**: Offers precise measurements of gene expression levels.
- **Discovery-Oriented**: Useful for discovering new genes, isoforms, and regulatory elements.

- **Lack of Single-Cell Resolution**: Bulk RNA-seq averages gene expression across all cells, potentially masking cell-specific variations.
- **Requires High-Quality RNA**: RNA degradation or contamination can affect results.
- **Complex Data Analysis**: Requires sophisticated computational tools and expertise for data processing and interpretation.

Key Steps in RNA-Seq

1. **Library Preparation**
   - Isolate RNA from the cells or tissues of interest
   - Convert RNA into complementary DNA (cDNA) and prepare sequencing libraries
2. **Sequencing technology** (e.g., Illumina, Oxford Nanopore)
3. **Data Analysis**

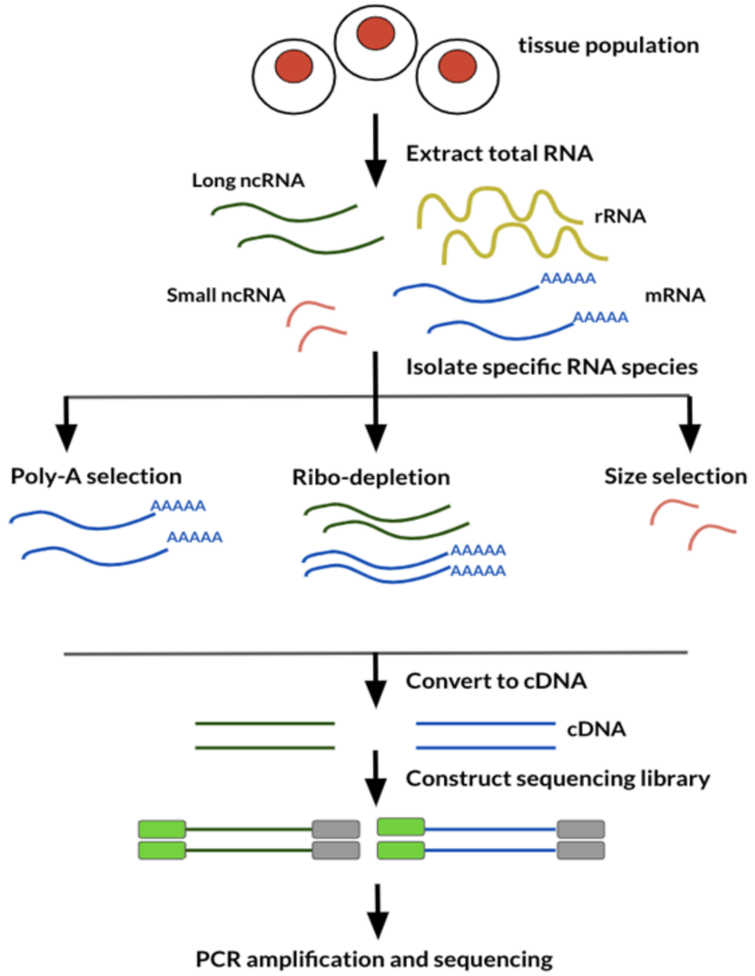
Alignment of reads against a reference genome or transcriptome

QC analysis

Data normalization

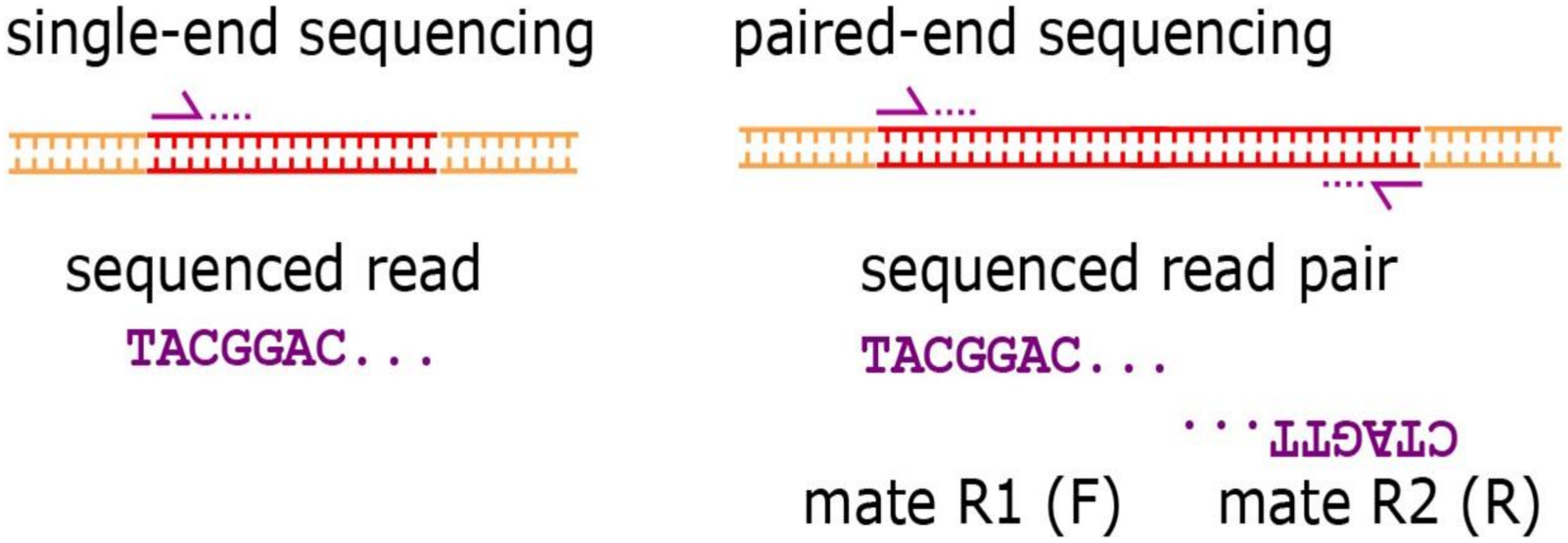
Quantify gene expression levels and analyze differential expression

How was the data produced?



Source: [Childhood Cancer Data Lab]

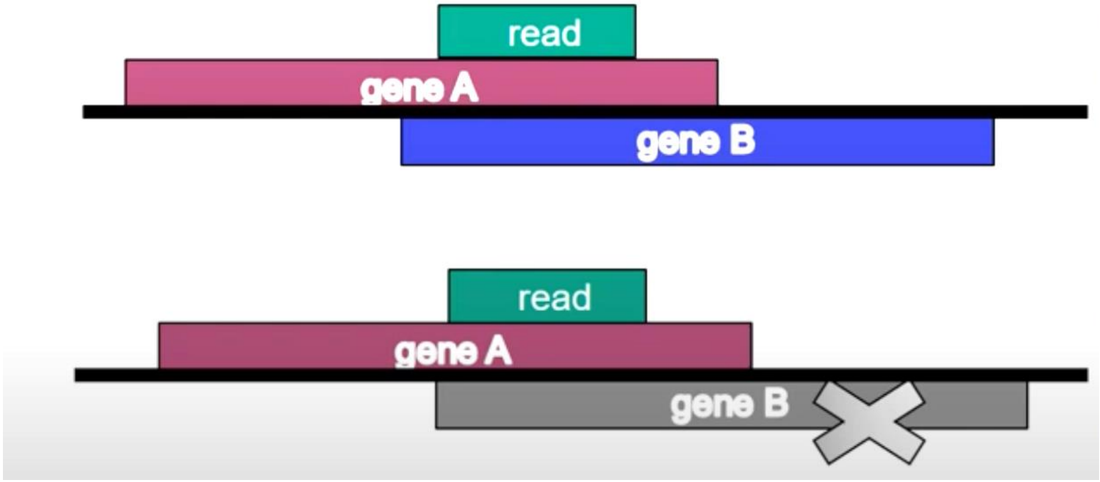
Single-end vs paired-end



Source: https://open.oregonstate.education/appliedbioinformatics/chapter/chapter-6/

Stranded RNA-Seq data

- Indicates if a read maps to same strand where the parental gene is, or to the opposite strand
  - Useful information when a read maps to a genomic location where there is a gene on both strands
- Several lab methods, you need to know which one was used
  - TruSeq stranded, NEB Ultra Directional, Agilent SureSelect Strand-Specific…



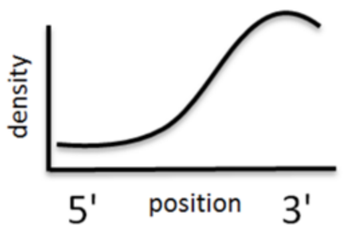
Unstranded data:
Does the read come
from geneA or geneB?

Stranded data:
the read comes from geneA

Source: https://cmb.molgen.mpg.de/2ndGenerationSequencing/Solas/RNA-seq.html

Sequence related biases

- **3’ bias -** 3’ ends of sequences are more likely to be sequenced

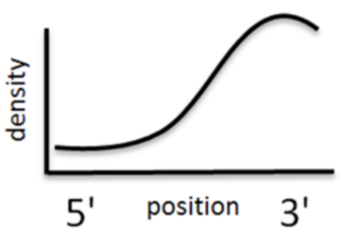


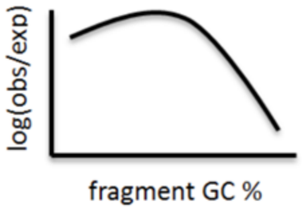
Source: `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

Sequence related biases

- **3’ bias -** 3’ ends of sequences are more likely to be sequenced

- **GC bias** - guanine and cytosine bond melts at higher temp - if a sequence has a lot of G’s and C’s

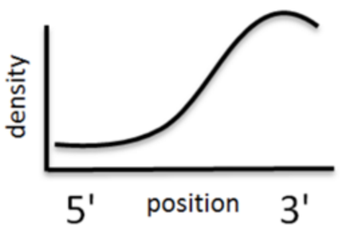




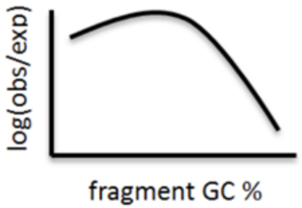
Source: `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

Sequence related biases

- **3’ bias -** 3’ ends of sequences are more likely to be sequenced



- **GC bias** - guanine and cytosine bond melts at higher temp - if a sequence has a lot of G’s and C’s

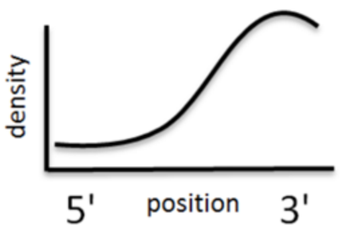


- **Sequence complexity** - certain sequences more likely to have primers bound to them (and more likely to be sequenced)

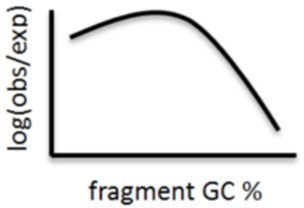
Source: `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

Sequence related biases

- **3’ bias -** 3’ ends of sequences are more likely to be sequenced

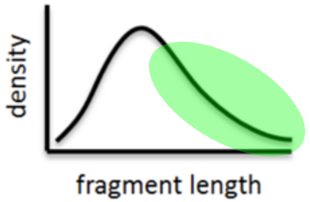


- **GC bias** - guanine and cytosine bond melts at higher temp - if a sequence has a lot of G’s and C’s



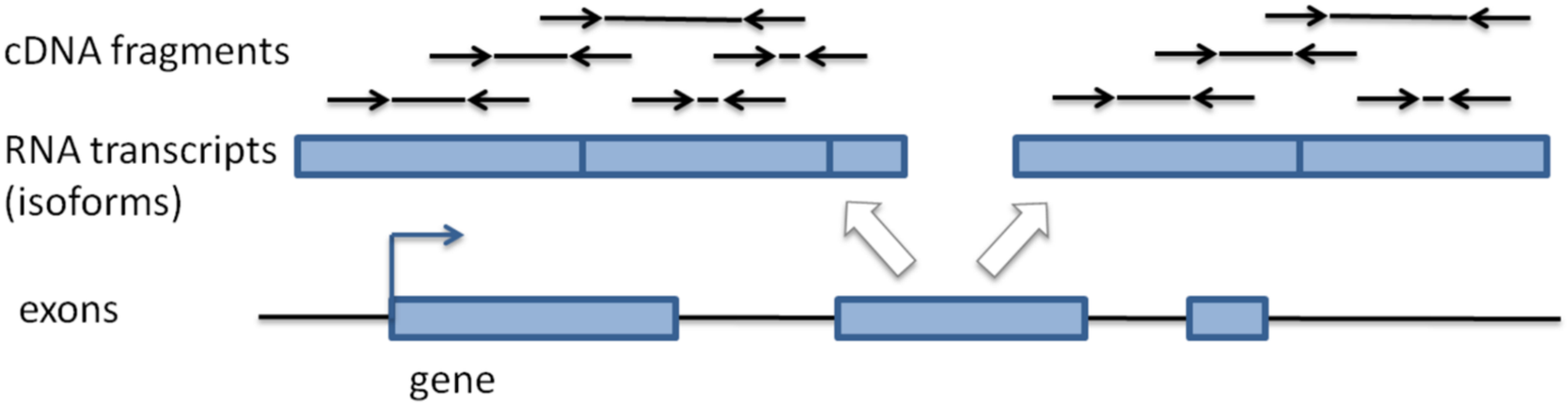
- **Sequence complexity** - certain sequences more likely to have primers bound to them (and more likely to be sequenced)

- **Length bias** - longer targets are more likely to be amplified or sequenced



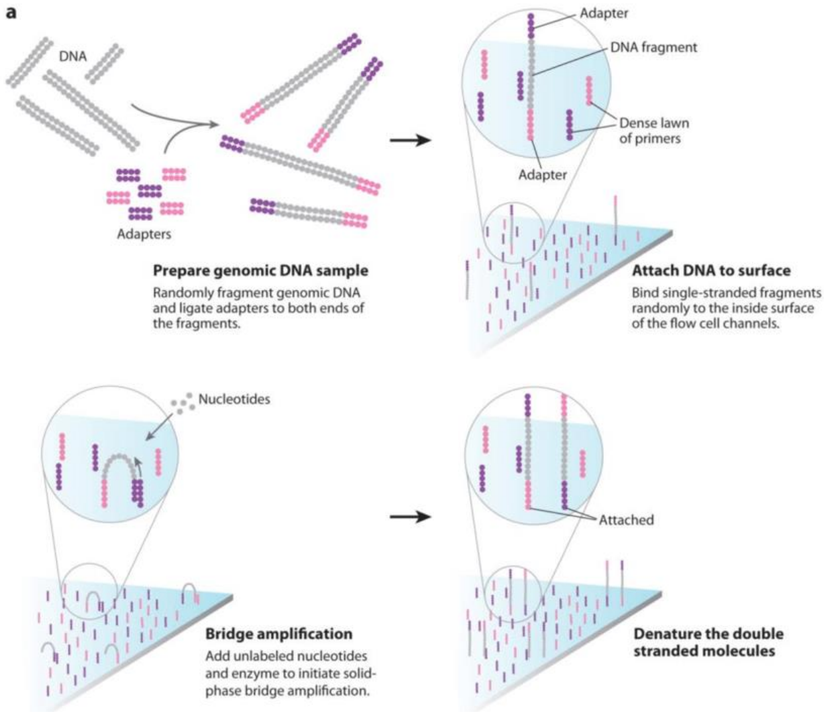
Source: `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

Sequence related biases: example



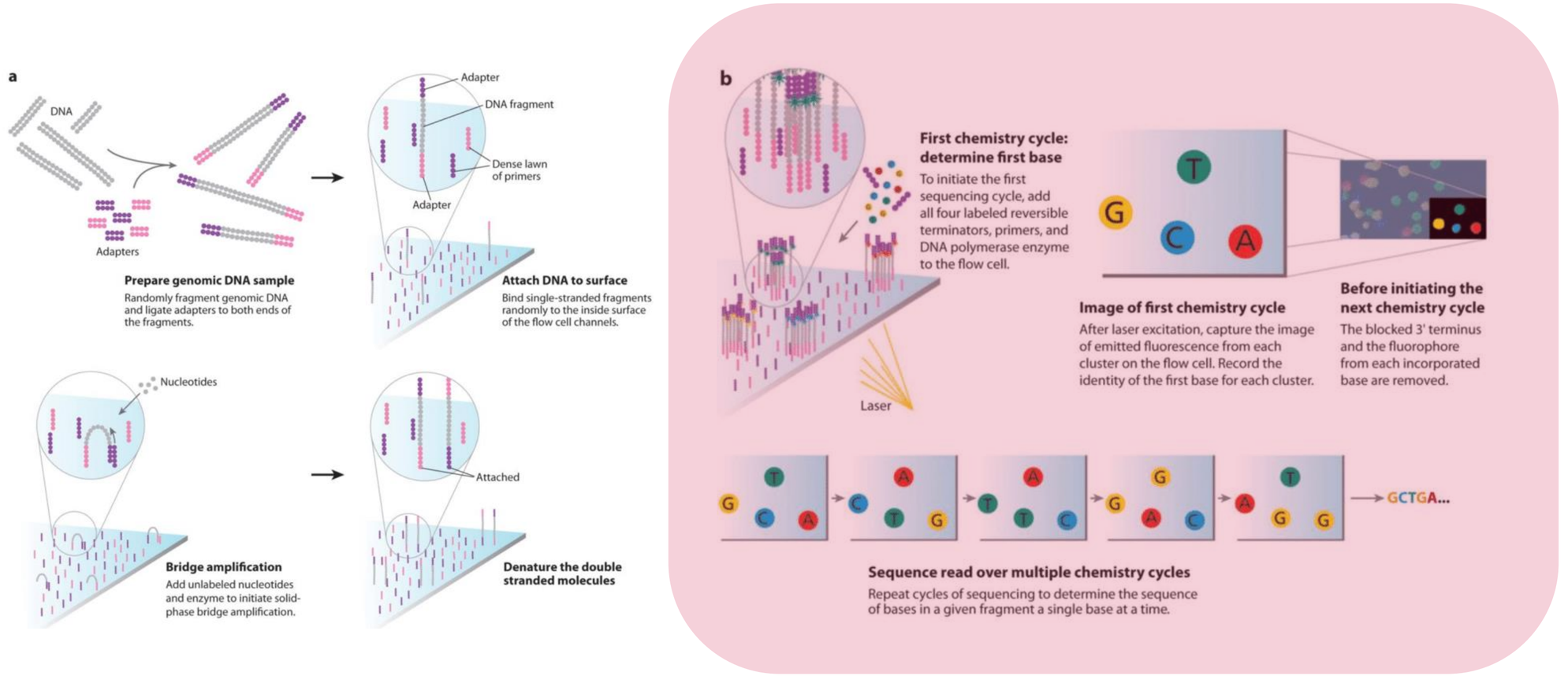
Source: Love *et al.* 2016, Nature Biotechnology.

RNA-Seq technology: Illumina sequencing (typically Hartwell)

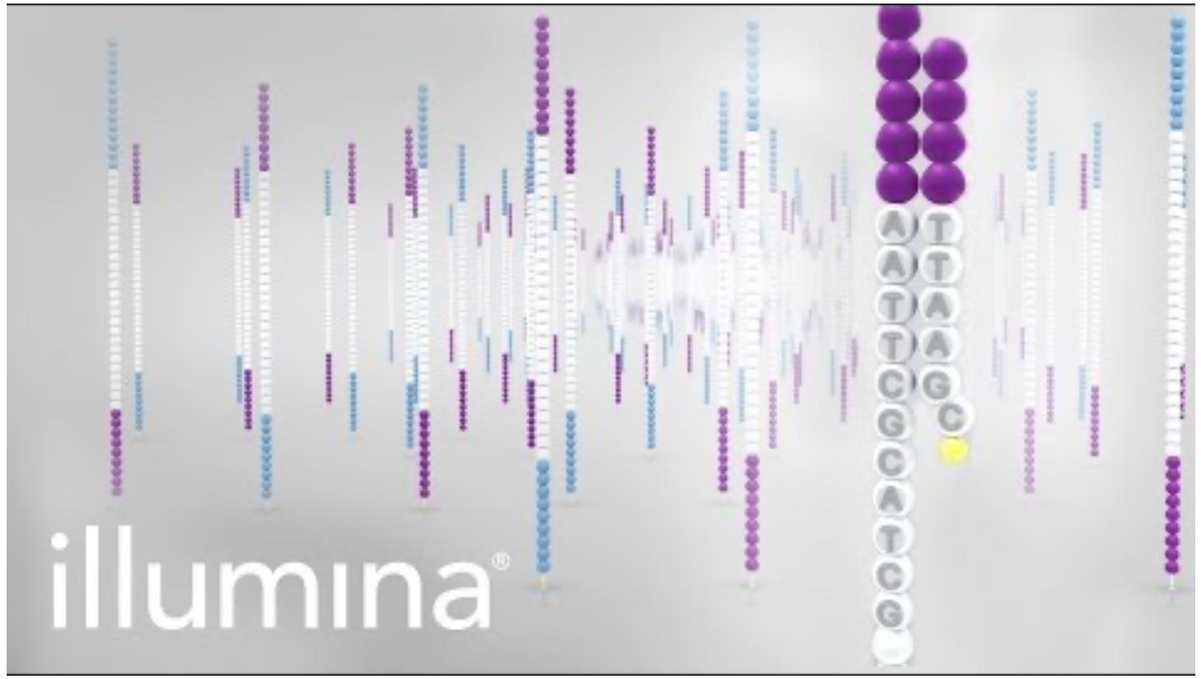


1. Sample Preparation
2. RNA Quality Assessment
3. Library Preparation
4. cDNA Synthesis
5. Fragmentation
6. Adapter Ligation
7. Library Amplification
8. PCR Amplification

RNA-Seq technology: Illumina sequencing (typically Hartwell)

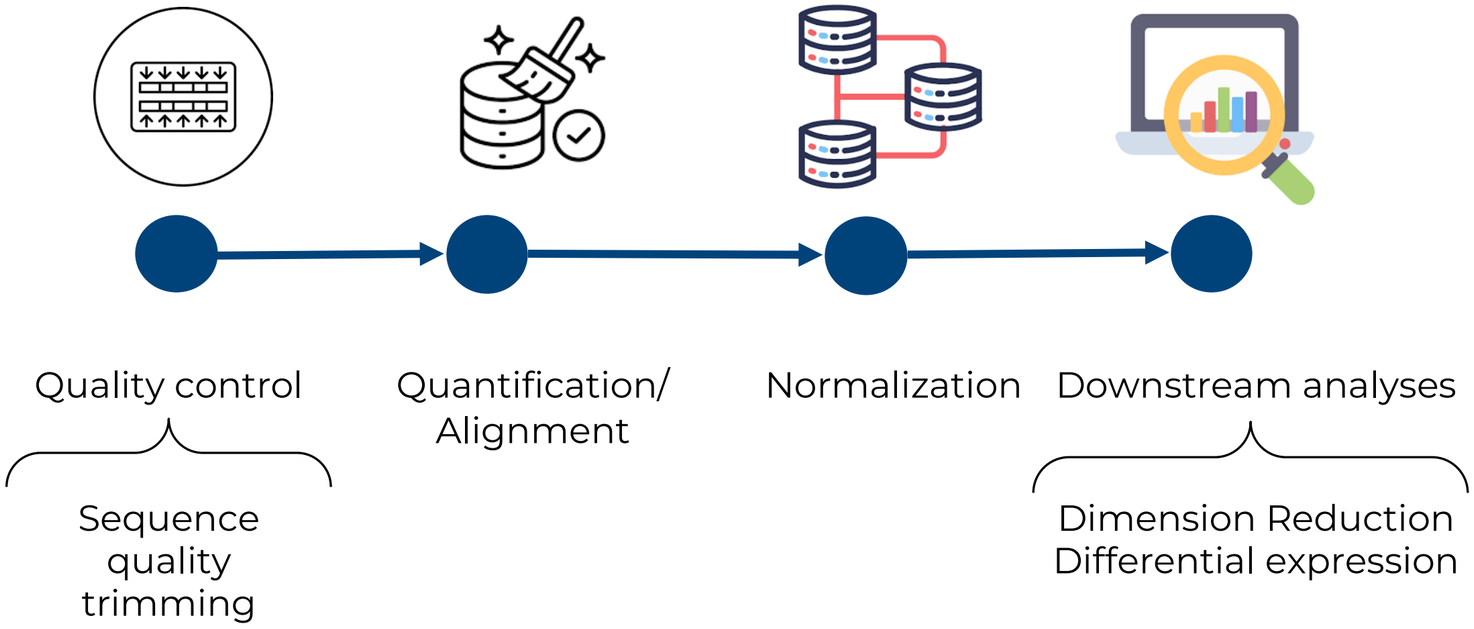


RNA-Seq technology: Illumina sequencing (typically Hartwell)



Overview of Illumina Sequencing by Synthesis Workflow

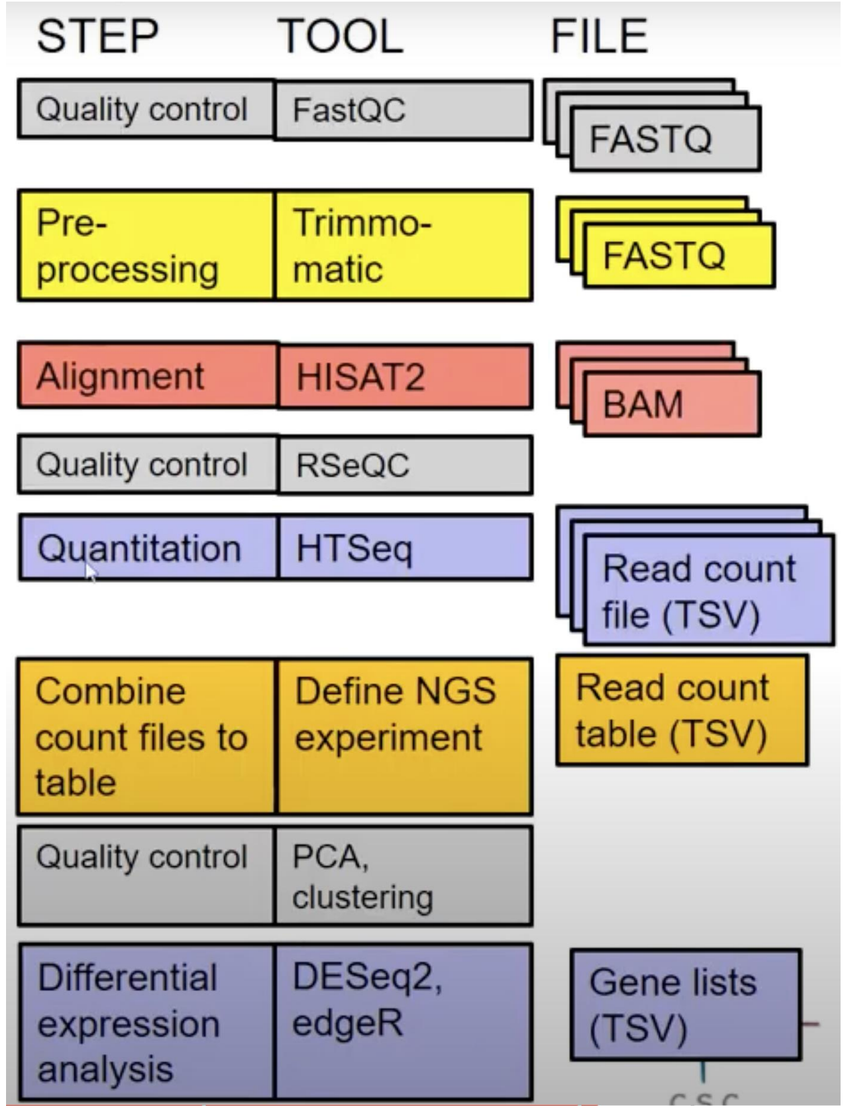
RNA-seq computational workflow



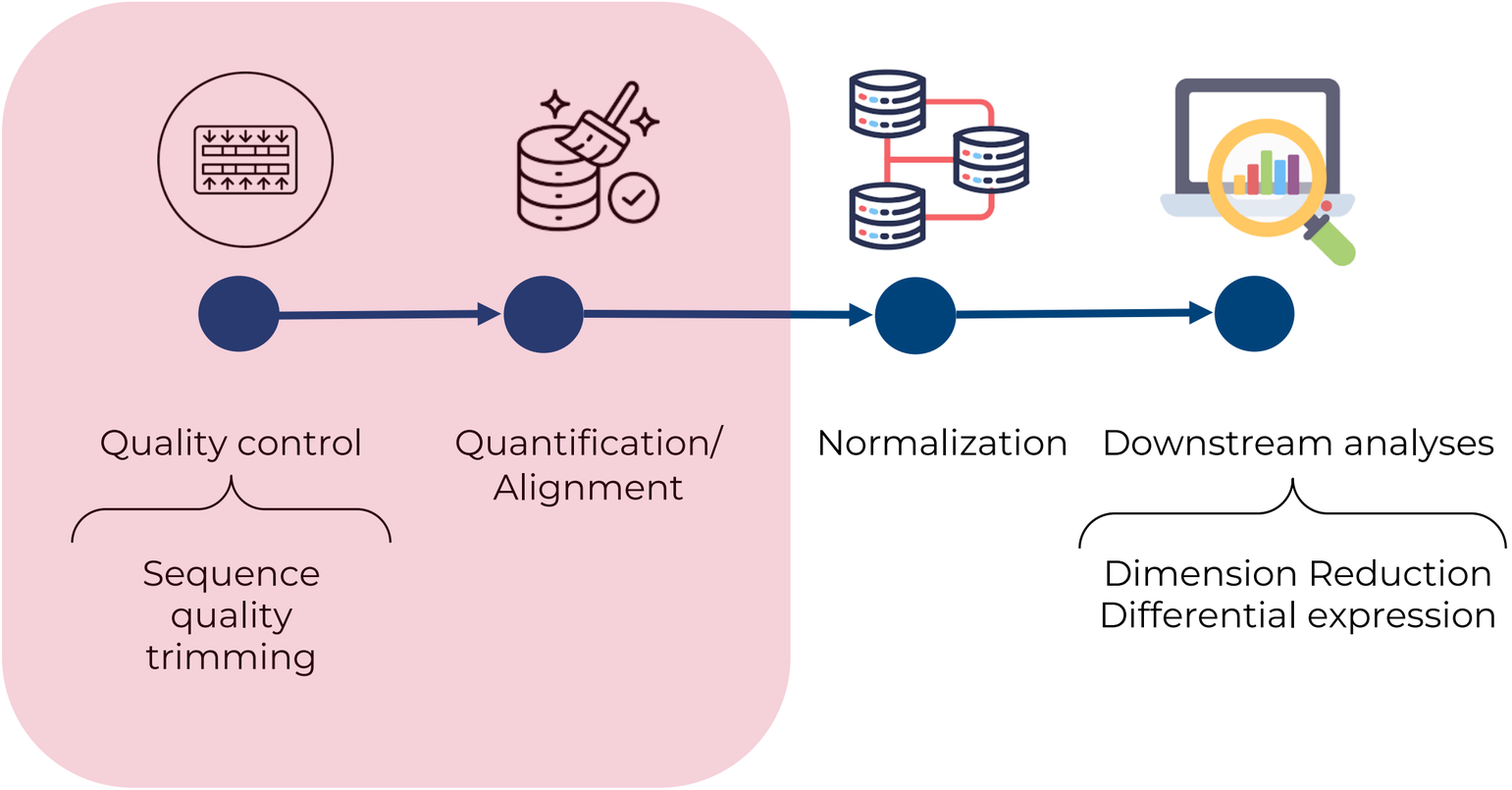
Source (edited): `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

RNA-seq computational workflow

So many steps, tools and files..

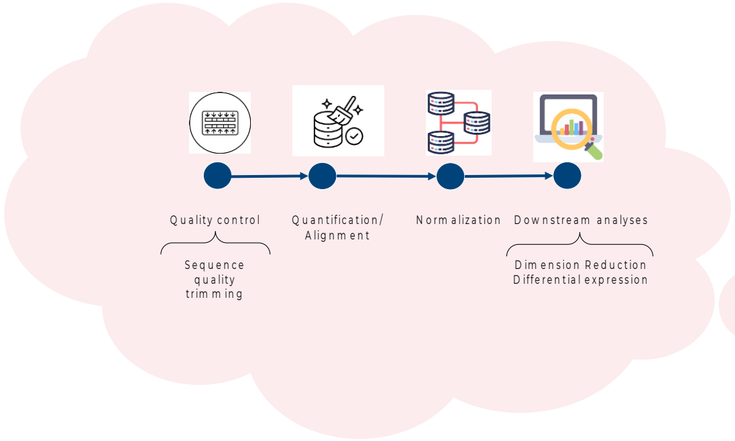


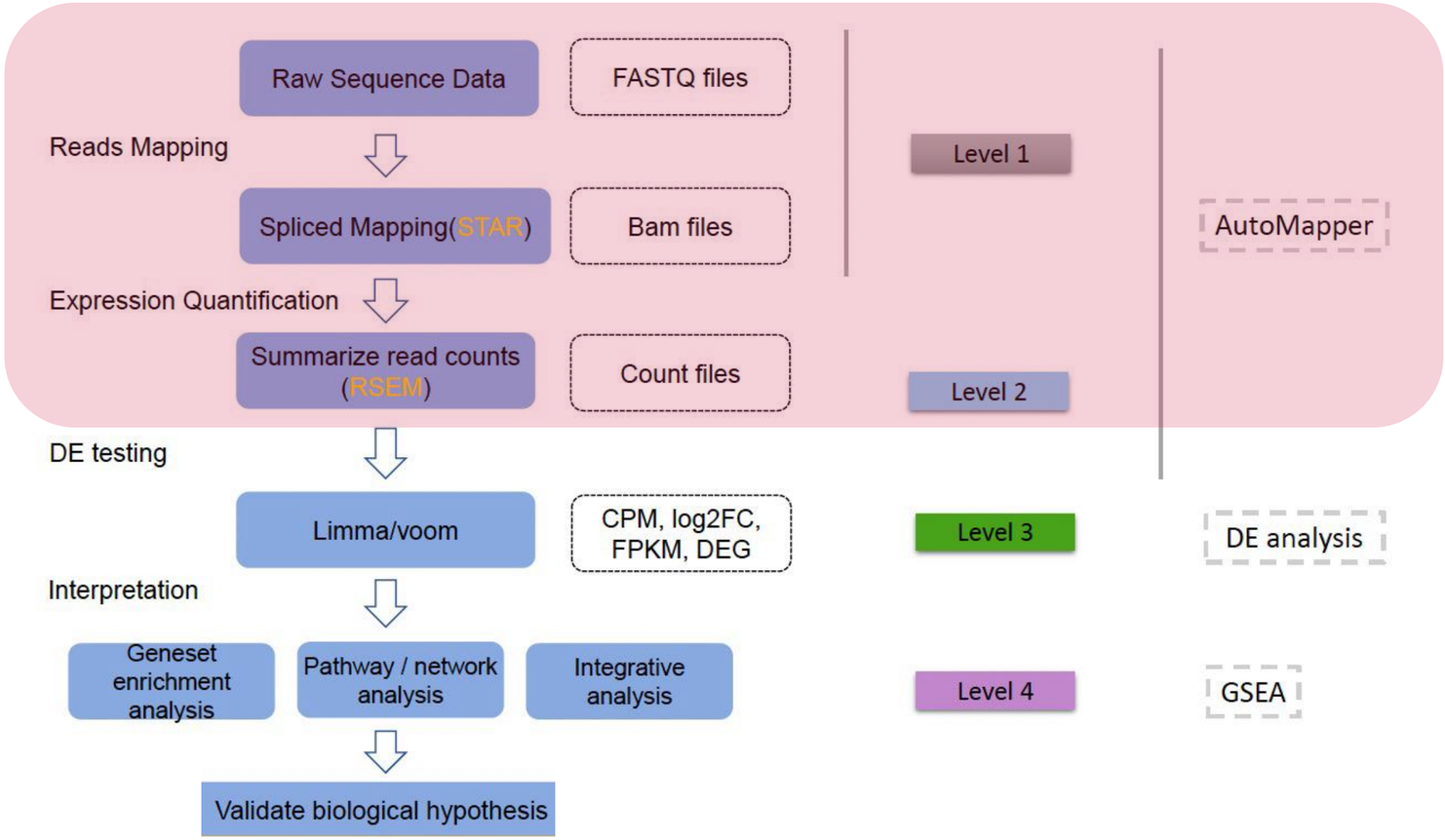
RNA-seq computational workflow



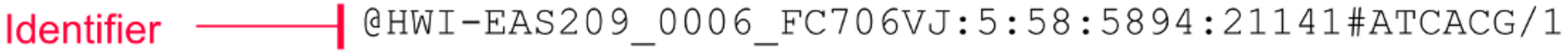
Source (edited): `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

Alignment and QC: CAB AutoMaper RNA-Seq pipeline



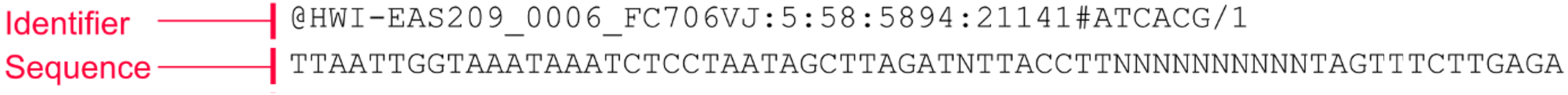


What is a FASTQ file?



Source: `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

What is a FASTQ file?



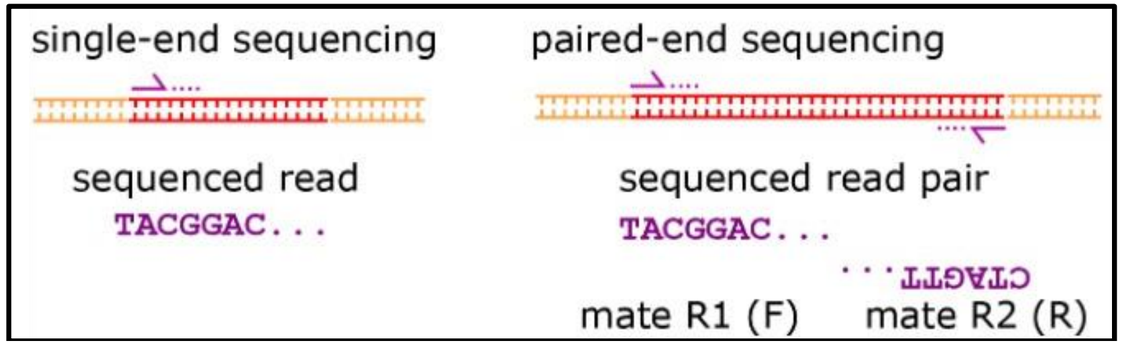
Source: `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

What is a FASTQ file?

Identifier ——— @HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1

Sequence ——— TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA

+ sign & identifier — +HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1



Source: `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

What is a FASTQ file?

Identifier — @HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1

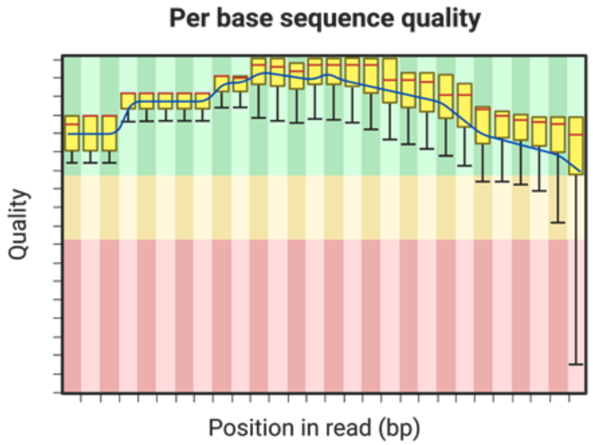
Sequence — TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA

+ sign & identifier — +HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1

Quality scores — efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`

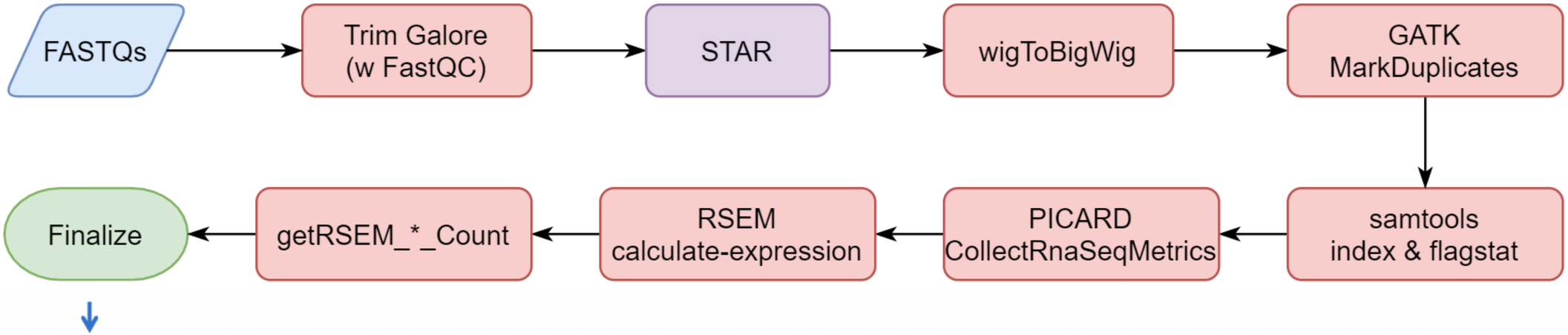
Base T

phred Quality] = 29



Source: `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

CAB AutoMaper RNA-Seq pipeline



.bam file
RSEM_count.txt
RSEM_FPKM.txt
RSEM_TPM.txt

SRM.report
SRM_*.txt

Most recent reference genomes:

- Hg38
- Mm10

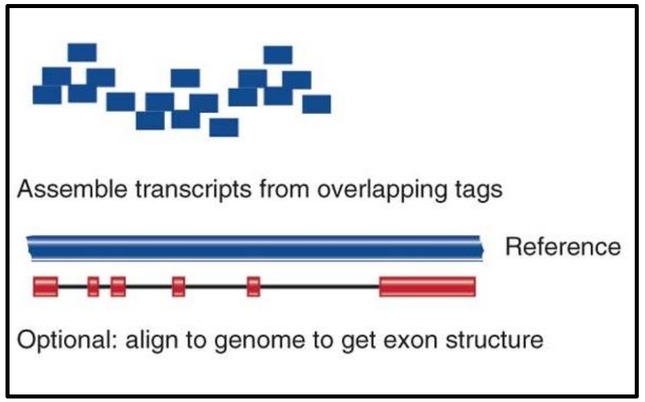
By Dr. Yawei Hui

https://wiki.stjude.org/display/CAB/AutoMapper#AutoMapper-E.RNA-sequencing(RNA-seq)

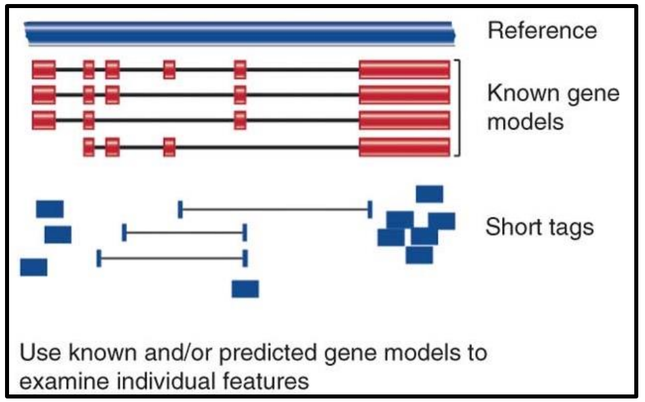
RNA-Seq mapping strategies

- RSEM (RNA-Seq by Expectation Maximization) is an accurate and user-friendly software tool for quantifying transcript abundances from RNA-Seq data.
- As it does not rely on the existence of a reference genome, it is particularly useful for quantification with de novo transcriptome assemblies.
- RSEM has enabled valuable guidance for cost-efficient design of quantification experiments with RNA-Seq, which is currently relatively expensive.

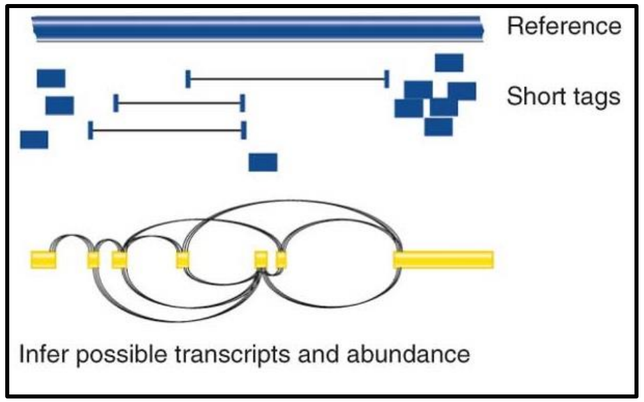
De novo assembly



Align to transcriptome



Align to reference genome

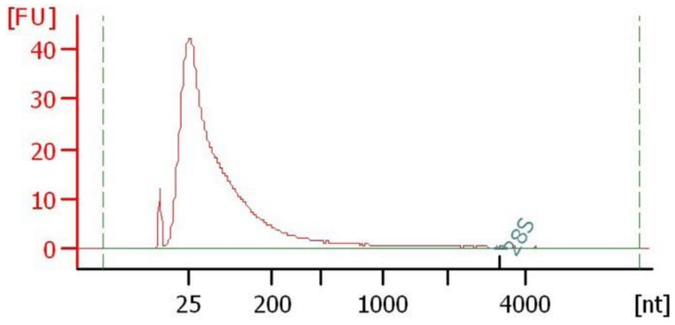


Diagrams from Cloonan & Grimmond, Nature Methods 2010

Quality assessment and Quality scores (QC)

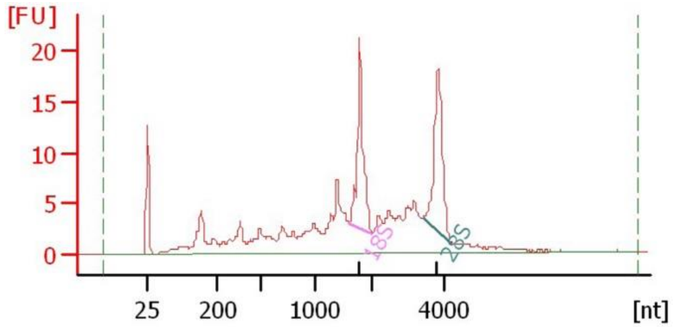
- **RNA Integrity Number (RIN) range: 0 – 10 (In lab)**

Trash quality



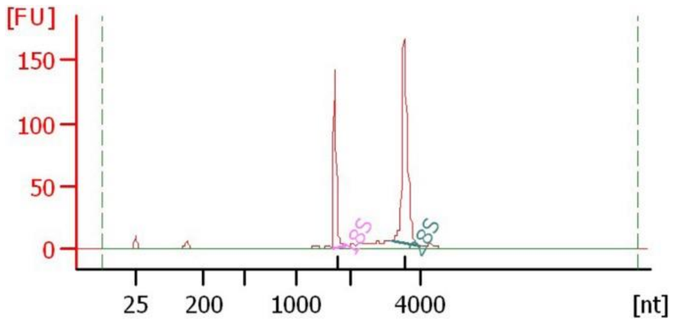
RIN = 1

Acceptable quality



RIN = 7

High quality



RIN = 10

Source: https://github.com/griffithlab/rnaseq_tutorial_wiki

CAB AutoMaper RNA-Seq pipeline: QC matrix

SAMPLE	GROUP	STATUS	RAW	TRIMMED	MAPPED	DUPLICATION	UNMAPPED
CONT-001	SKNAS-CONT	PASS	295047824	294129654	269846318	39.91%	8.25%
CONT-002	SKNAS-CONT	PASS	245531910	244358138	233188386	43.12%	4.56%
CONT-003	SKNAS-CONT	PASS	253659252	252818094	234745872	50.33%	7.13%
TUMOR-001	SKNAS-TUMOR	PASS	262184424	261302964	241432954	36.07%	7.59%
TUMOR-002	SKNAS-TUMOR	PASS	267545786	266472740	240811496	40.79%	9.62%
TUMOR-003	SKNAS-TUMOR	PASS	242060090	240936958	233931642	58.56%	2.9%

SAMPLE	GROUP	STATUS	MAPPED	CODING	UTR	INTRON	INTERGENIC	RIBOSOMAL
CONT-001	SKNAS-CONT	PASS	91.74%	45.93%	38.99%	9.69%	5.32%	0.07%
CONT-002	SKNAS-CONT	PASS	95.43%	47.46%	38.79%	7.75%	5.91%	0.09%
CONT-003	SKNAS-CONT	PASS	92.85%	34.27%	43.45%	14.34%	7.78%	0.16%
TUMOR-001	SKNAS-TUMOR	PASS	92.4%	40.51%	36.77%	14.82%	7.81%	0.11%
TUMOR-002	SKNAS-TUMOR	PASS	90.37%	38.99%	36.19%	16.96%	7.78%	0.07%
TUMOR-003	SKNAS-TUMOR	PASS	97.09%	29.94%	50.57%	12.58%	6.83%	0.09%

CAB AutoMaper RNA-Seq pipeline: QC matrix

- RAW: Reads included in the original FASTQs

SAMPLE	GROUP	STATUS	RAW	TRIMMED	MAPPED	DUPLICATION	UNMAPPED
CONT-001	SKNAS-CONT	PASS	295047824	294129654	269846318	39.91%	8.25%
CONT-002	SKNAS-CONT	PASS	245531910	244358138	233188386	43.12%	4.56%
CONT-003	SKNAS-CONT	PASS	253659252	252818094	234745872	50.33%	7.13%
TUMOR-001	SKNAS-TUMOR	PASS	262184424	261302964	241432954	36.07%	7.59%
TUMOR-002	SKNAS-TUMOR	PASS	267545786	266472740	240811496	40.79%	9.62%
TUMOR-003	SKNAS-TUMOR	PASS	242060090	240936958	233931642	58.56%	2.9%

SAMPLE	GROUP	STATUS	MAPPED	CODING	UTR	INTRON	INTERGENIC	RIBOSOMAL
CONT-001	SKNAS-CONT	PASS	91.74%	45.93%	38.99%	9.69%	5.32%	0.07%
CONT-002	SKNAS-CONT	PASS	95.43%	47.46%	38.79%	7.75%	5.91%	0.09%
CONT-003	SKNAS-CONT	PASS	92.85%	34.27%	43.45%	14.34%	7.78%	0.16%
TUMOR-001	SKNAS-TUMOR	PASS	92.4%	40.51%	36.77%	14.82%	7.81%	0.11%
TUMOR-002	SKNAS-TUMOR	PASS	90.37%	38.99%	36.19%	16.96%	7.78%	0.07%
TUMOR-003	SKNAS-TUMOR	PASS	97.09%	29.94%	50.57%	12.58%	6.83%	0.09%

CAB AutoMaper RNA-Seq pipeline: QC matrix

- TRIMMED: Reads extracted from the raw reads after adapter trimming by using "trim-galore"

SAMPLE	GROUP	STATUS	RAW	TRIMMED	MAPPED	DUPLICATION	UNMAPPED
CONT-001	SKNAS-CONT	PASS	295047824	294129654	269846318	39.91%	8.25%
CONT-002	SKNAS-CONT	PASS	245531910	244358138	233188386	43.12%	4.56%
CONT-003	SKNAS-CONT	PASS	253659252	252818094	234745872	50.33%	7.13%
TUMOR-001	SKNAS-TUMOR	PASS	262184424	261302964	241432954	36.07%	7.59%
TUMOR-002	SKNAS-TUMOR	PASS	267545786	266472740	240811496	40.79%	9.62%
TUMOR-003	SKNAS-TUMOR	PASS	242060090	240936958	233931642	58.56%	2.9%

SAMPLE	GROUP	STATUS	MAPPED	CODING	UTR	INTRON	INTERGENIC	RIBOSOMAL
CONT-001	SKNAS-CONT	PASS	91.74%	45.93%	38.99%	9.69%	5.32%	0.07%
CONT-002	SKNAS-CONT	PASS	95.43%	47.46%	38.79%	7.75%	5.91%	0.09%
CONT-003	SKNAS-CONT	PASS	92.85%	34.27%	43.45%	14.34%	7.78%	0.16%
TUMOR-001	SKNAS-TUMOR	PASS	92.4%	40.51%	36.77%	14.82%	7.81%	0.11%
TUMOR-002	SKNAS-TUMOR	PASS	90.37%	38.99%	36.19%	16.96%	7.78%	0.07%
TUMOR-003	SKNAS-TUMOR	PASS	97.09%	29.94%	50.57%	12.58%	6.83%	0.09%

CAB AutoMaper RNA-Seq pipeline: QC matrix

- MAPPED: Reads include uniquely and multiple mapped reads while excluding those mapped to too many loci

SAMPLE	GROUP	STATUS	RAW	TRIMMED	MAPPED	DUPLICATION	UNMAPPED
CONT-001	SKNAS-CONT	PASS	295047824	294129654	269846318	39.91%	8.25%
CONT-002	SKNAS-CONT	PASS	245531910	244358138	233188386	43.12%	4.56%
CONT-003	SKNAS-CONT	PASS	253659252	252818094	234745872	50.33%	7.13%
TUMOR-001	SKNAS-TUMOR	PASS	262184424	261302964	241432954	36.07%	7.59%
TUMOR-002	SKNAS-TUMOR	PASS	267545786	266472740	240811496	40.79%	9.62%
TUMOR-003	SKNAS-TUMOR	PASS	242060090	240936958	233931642	58.56%	2.9%

SAMPLE	GROUP	STATUS	MAPPED	CODING	UTR	INTRON	INTERGENIC	RIBOSOMAL
CONT-001	SKNAS-CONT	PASS	91.74%	45.93%	38.99%	9.69%	5.32%	0.07%
CONT-002	SKNAS-CONT	PASS	95.43%	47.46%	38.79%	7.75%	5.91%	0.09%
CONT-003	SKNAS-CONT	PASS	92.85%	34.27%	43.45%	14.34%	7.78%	0.16%
TUMOR-001	SKNAS-TUMOR	PASS	92.4%	40.51%	36.77%	14.82%	7.81%	0.11%
TUMOR-002	SKNAS-TUMOR	PASS	90.37%	38.99%	36.19%	16.96%	7.78%	0.07%
TUMOR-003	SKNAS-TUMOR	PASS	97.09%	29.94%	50.57%	12.58%	6.83%	0.09%

CAB AutoMaper RNA-Seq pipeline: QC matrix

- UNMAPPED: Reads include unmapped reads due to 1) too many mismatches; 2) too short; 3) all other

SAMPLE	GROUP	STATUS	RAW	TRIMMED	MAPPED	DUPLICATION	UNMAPPED
CONT-001	SKNAS-CONT	PASS	295047824	294129654	269846318	39.91%	8.25%
CONT-002	SKNAS-CONT	PASS	245531910	244358138	233188386	43.12%	4.56%
CONT-003	SKNAS-CONT	PASS	253659252	252818094	234745872	50.33%	7.13%
TUMOR-001	SKNAS-TUMOR	PASS	262184424	261302964	241432954	36.07%	7.59%
TUMOR-002	SKNAS-TUMOR	PASS	267545786	266472740	240811496	40.79%	9.62%
TUMOR-003	SKNAS-TUMOR	PASS	242060090	240936958	233931642	58.56%	2.9%

SAMPLE	GROUP	STATUS	MAPPED	CODING	UTR	INTRON	INTERGENIC	RIBOSOMAL
CONT-001	SKNAS-CONT	PASS	91.74%	45.93%	38.99%	9.69%	5.32%	0.07%
CONT-002	SKNAS-CONT	PASS	95.43%	47.46%	38.79%	7.75%	5.91%	0.09%
CONT-003	SKNAS-CONT	PASS	92.85%	34.27%	43.45%	14.34%	7.78%	0.16%
TUMOR-001	SKNAS-TUMOR	PASS	92.4%	40.51%	36.77%	14.82%	7.81%	0.11%
TUMOR-002	SKNAS-TUMOR	PASS	90.37%	38.99%	36.19%	16.96%	7.78%	0.07%
TUMOR-003	SKNAS-TUMOR	PASS	97.09%	29.94%	50.57%	12.58%	6.83%	0.09%

CAB AutoMaper RNA-Seq pipeline: QC matrix

- ATTENTION: Starting from Nov. 22nd, 2020, the RNAseq mapping pipeline is updated to deliver a BAM file with duplicate reads marked by GATK, after it is generated by STAR.

SAMPLE	GROUP	STATUS	RAW	TRIMMED	MAPPED	DUPLICATION	UNMAPPED
CONT-001	SKNAS-CONT	PASS	295047824	294129654	269846318	39.91%	8.25%
CONT-002	SKNAS-CONT	PASS	245531910	244358138	233188386	43.12%	4.56%
CONT-003	SKNAS-CONT	PASS	253659252	252818094	234745872	50.33%	7.13%
TUMOR-001	SKNAS-TUMOR	PASS	262184424	261302964	241432954	36.07%	7.59%
TUMOR-002	SKNAS-TUMOR	PASS	267545786	266472740	240811496	40.79%	9.62%
TUMOR-003	SKNAS-TUMOR	PASS	242060090	240936958	233931642	58.56%	2.9%

SAMPLE	GROUP	STATUS	MAPPED	CODING	UTR	INTRON	INTERGENIC	RIBOSOMAL
CONT-001	SKNAS-CONT	PASS	91.74%	45.93%	38.99%	9.69%	5.32%	0.07%
CONT-002	SKNAS-CONT	PASS	95.43%	47.46%	38.79%	7.75%	5.91%	0.09%
CONT-003	SKNAS-CONT	PASS	92.85%	34.27%	43.45%	14.34%	7.78%	0.16%
TUMOR-001	SKNAS-TUMOR	PASS	92.4%	40.51%	36.77%	14.82%	7.81%	0.11%
TUMOR-002	SKNAS-TUMOR	PASS	90.37%	38.99%	36.19%	16.96%	7.78%	0.07%
TUMOR-003	SKNAS-TUMOR	PASS	97.09%	29.94%	50.57%	12.58%	6.83%	0.09%

CAB AutoMaper RNA-Seq pipeline: QC matrix

- Status: PASS/WARNING

SAMPLE	GROUP	STATUS	RAW	TRIMMED	MAPPED	DUPLICATION	UNMAPPED
CONT-001	SKNAS-CONT	PASS	295047824	294129654	269846318	39.91%	8.25%
CONT-002	SKNAS-CONT	PASS	245531910	244358138	233188386	43.12%	4.56%
CONT-003	SKNAS-CONT	PASS	253659252	252818094	234745872	50.33%	7.13%
TUMOR-001	SKNAS-TUMOR	PASS	262184424	261302964	241432954	36.07%	7.59%
TUMOR-002	SKNAS-TUMOR	PASS	267545786	266472740	240811496	40.79%	9.62%
TUMOR-003	SKNAS-TUMOR	PASS	242060090	240936958	233931642	58.56%	2.9%

SAMPLE	GROUP	STATUS	MAPPED	CODING	UTR	INTRON	INTERGENIC	RIBOSOMAL
CONT-001	SKNAS-CONT	PASS	91.74%	45.93%	38.99%	9.69%	5.32%	0.07%
CONT-002	SKNAS-CONT	PASS	95.43%	47.46%	38.79%	7.75%	5.91%	0.09%
CONT-003	SKNAS-CONT	PASS	92.85%	34.27%	43.45%	14.34%	7.78%	0.16%
TUMOR-001	SKNAS-TUMOR	PASS	92.4%	40.51%	36.77%	14.82%	7.81%	0.11%
TUMOR-002	SKNAS-TUMOR	PASS	90.37%	38.99%	36.19%	16.96%	7.78%	0.07%
TUMOR-003	SKNAS-TUMOR	PASS	97.09%	29.94%	50.57%	12.58%	6.83%	0.09%

CAB AutoMaper RNA-Seq pipeline: QC matrix

		Raw	Trimmed	Mapped	Unmapped
57	PASS	183,713,112	182,843,098	174,654,868	8,071,166
58	PASS	206,911,576	206,567,604	200,732,364	5,705,280
59	PASS	179,434,874	179,063,722	173,428,776	5,535,678
60	PASS	175,251,394	174,982,466	169,957,226	4,923,034
61	PASS	188,300,318	187,985,030	182,120,406	5,762,370
METRICS	WARNING	< 90M		< 60M	
	PASS	≥ 90M		≥ 60M	

CAB AutoMaper RNA-Seq pipeline: QC matrix

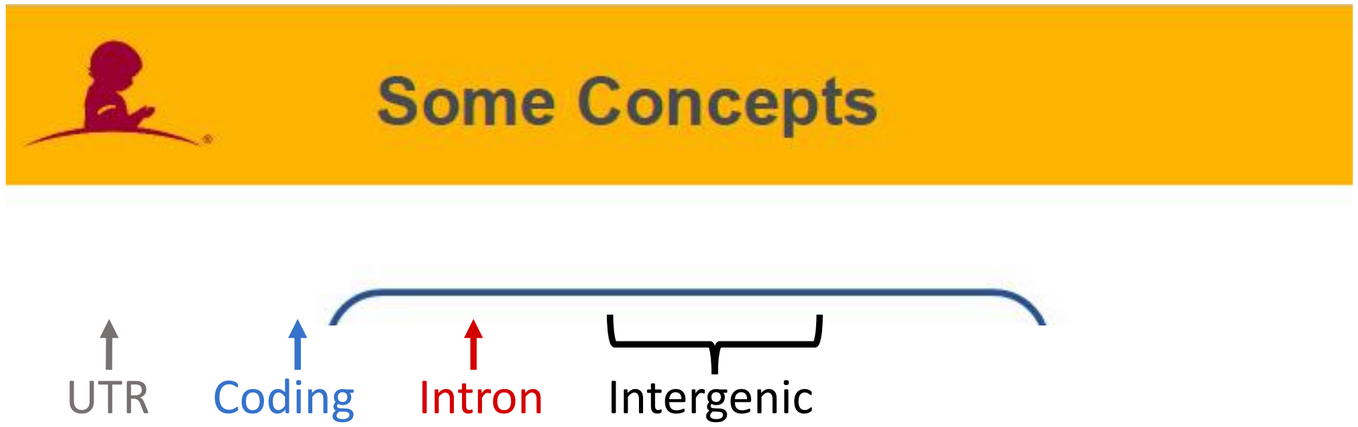
- Read Alignment and Distribution

SAMPLE	GROUP	STATUS	RAW	TRIMMED	MAPPED	DUPLICATION	UNMAPPED
CONT-001	SKNAS-CONT	PASS	295047824	294129654	269846318	39.91%	8.25%
CONT-002	SKNAS-CONT	PASS	245531910	244358138	233188386	43.12%	4.56%
CONT-003	SKNAS-CONT	PASS	253659252	252818094	234745872	50.33%	7.13%
TUMOR-001	SKNAS-TUMOR	PASS	262184424	261302964	241432954	36.07%	7.59%
TUMOR-002	SKNAS-TUMOR	PASS	267545786	266472740	240811496	40.79%	9.62%
TUMOR-003	SKNAS-TUMOR	PASS	242060090	240936958	233931642	58.56%	2.9%

SAMPLE	GROUP	STATUS	MAPPED	CODING	UTR	INTRON	INTERGENIC	RIBOSOMAL
CONT-001	SKNAS-CONT	PASS	91.74%	45.93%	38.99%	9.69%	5.32%	0.07%
CONT-002	SKNAS-CONT	PASS	95.43%	47.46%	38.79%	7.75%	5.91%	0.09%
CONT-003	SKNAS-CONT	PASS	92.85%	34.27%	43.45%	14.34%	7.78%	0.16%
TUMOR-001	SKNAS-TUMOR	PASS	92.4%	40.51%	36.77%	14.82%	7.81%	0.11%
TUMOR-002	SKNAS-TUMOR	PASS	90.37%	38.99%	36.19%	16.96%	7.78%	0.07%
TUMOR-003	SKNAS-TUMOR	PASS	97.09%	29.94%	50.57%	12.58%	6.83%	0.09%

Read Alignment and Distribution

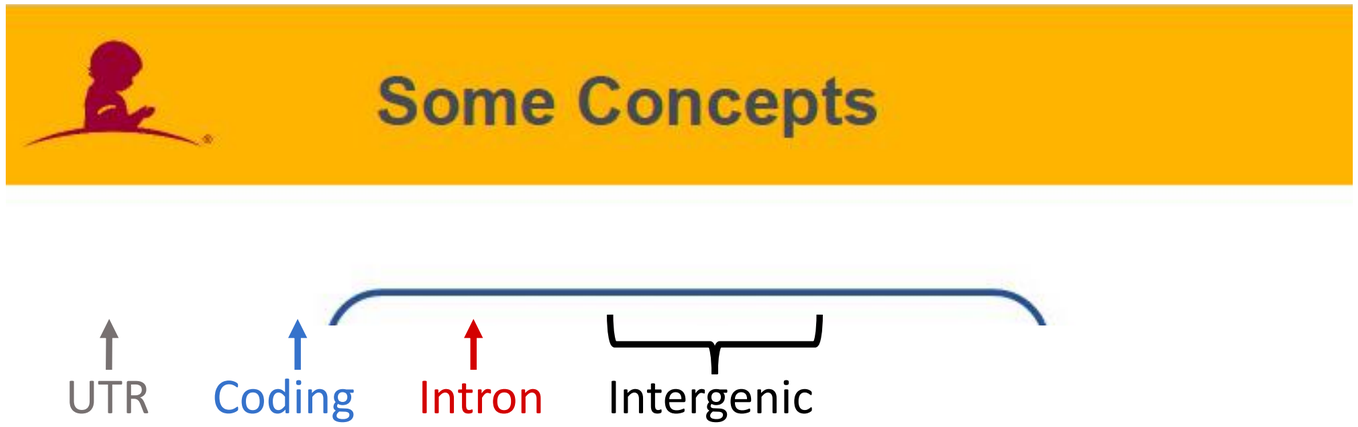
- Majority of reads to be in exons
- A large number of intronic/intergenic reads can indicate DNA contamination
- But a high number of intronic reads may be expected for total RNA preparations as pre-mRNAs are not explicitly excluded.
- DNA contamination is usually apparent by eye when viewing in IGV, as many intergenic reads will be apparent.
- Nascent transcription (which brain tissue has a lot of) and high levels of transcriptional activity can contribute to a high number of intronic reads. As such, these ranges can vary quite a bit by tissue/cell type and library preparation methods.



Read Alignment and Distribution

Typical ranges for PolyA selected libraries

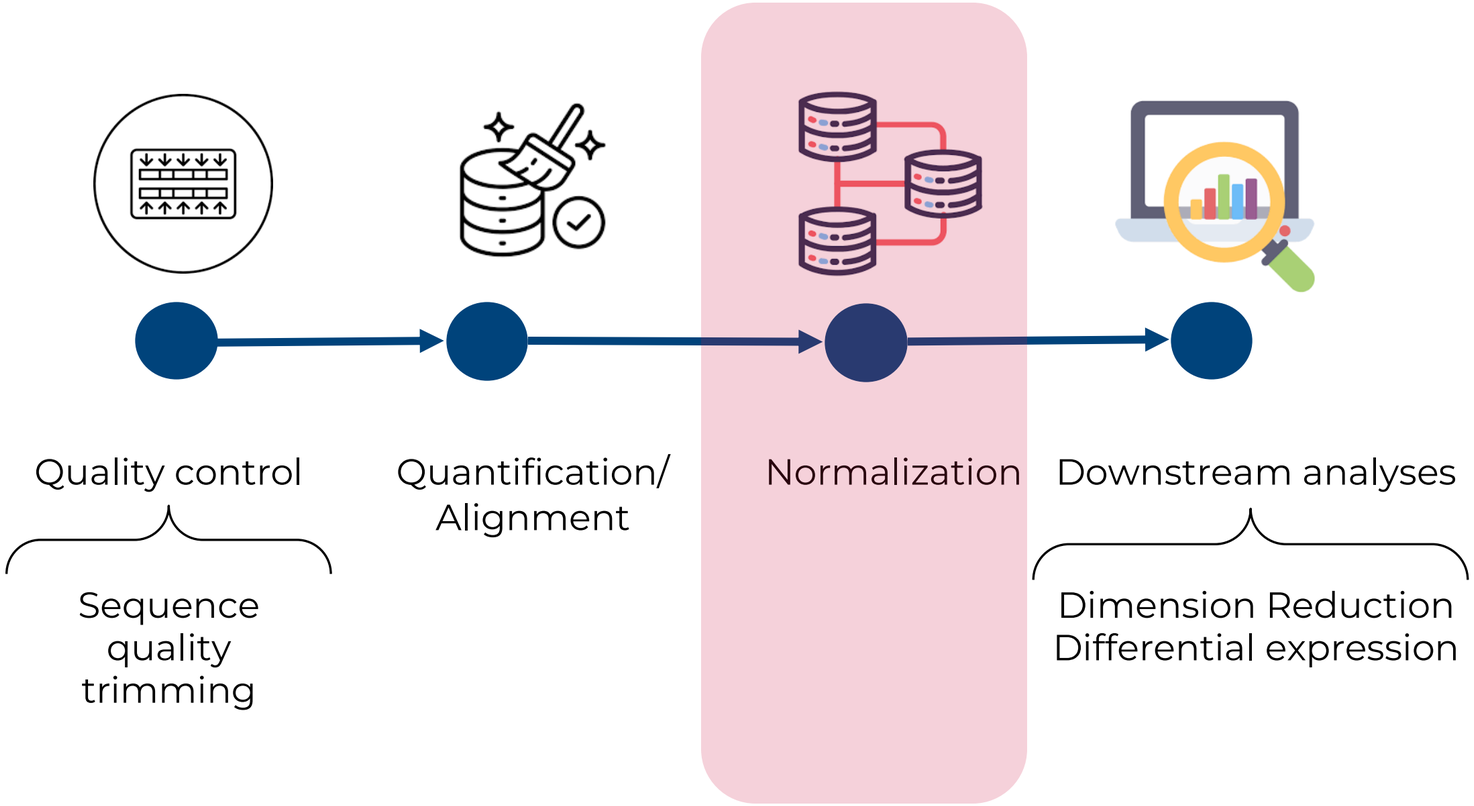
- Exonic reads: 60-90%
- Intronic reads: Up to 30%
- Intergenic reads: <10%



Read Alignment and Distribution

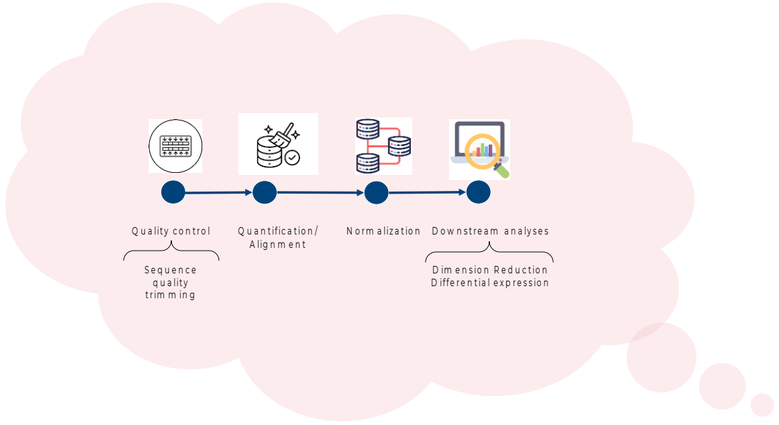
	Status								
					Exonic		Intron	Intergenic	Ribosomal
		Mapped	Unmapped	Duplication	Coding	UTR			
57	PASS	95.52	4.41	43.69	26.00	20.12	32.12	20.91	0.85
58	PASS	97.18	2.76	39.98	24.06	19.15	35.74	20.30	0.75
59	PASS	96.85	3.09	44.82	29.18	22.54	23.01	24.14	1.14
60	PASS	97.13	2.81	42.16	25.71	20.14	31.88	21.37	0.90
61	PASS	96.88	3.07	45.22	31.07	24.31	20.42	23.08	1.12
METRICS	WARNING	< 85	> 5	> 40	< 60		> 30	> 10	> 15
	PASS	≥ 85	≤ 5	≤ 40	≥ 60		≤ 30	≤ 10	≤ 15

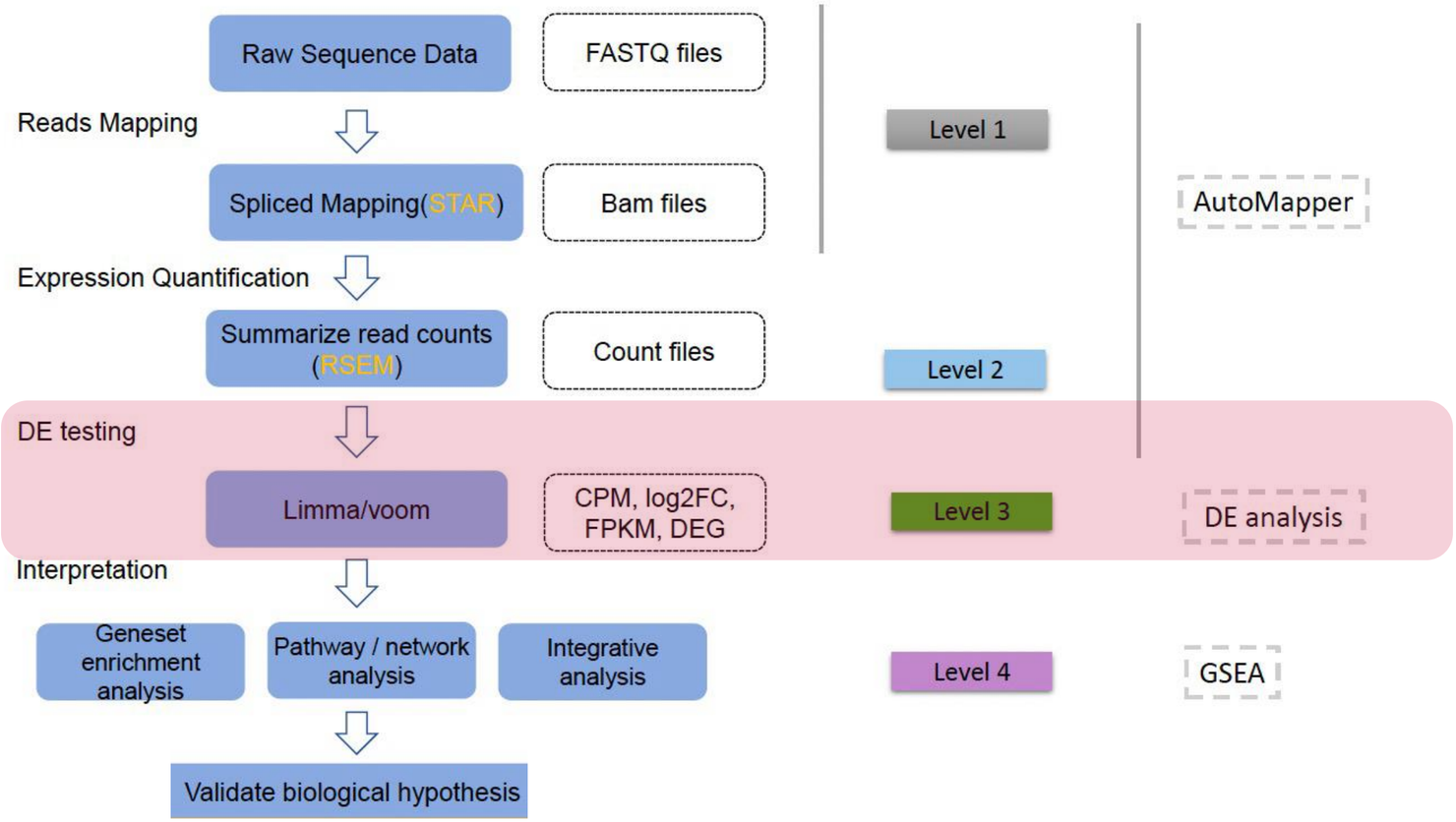
RNA-seq computational workflow



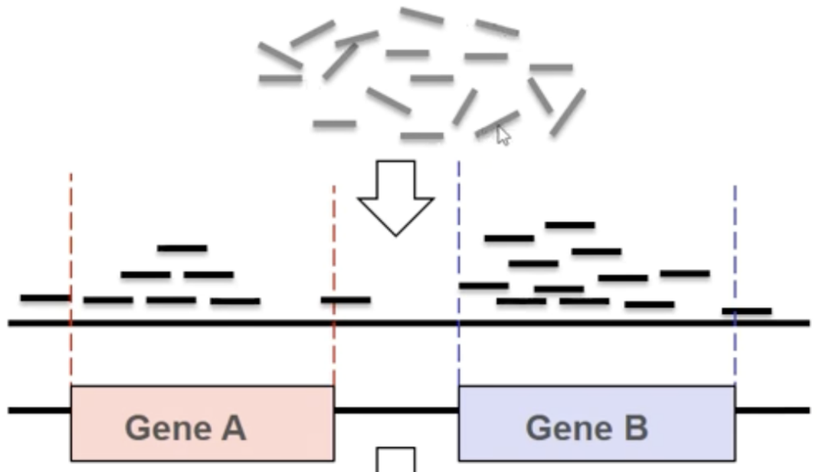
Source (edited): `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

Data normalization





Demo data



Raw data (reads)

Align reads to reference genome

Match alignment positions with known gene positions

Count how many reads each gene has

	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3
Gene A	6	5	7	170	100	110
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1

Compare sample groups: differential expression analysis

Data normalization

- Each sample will have different number of reads assigned to it due to the fact that one sample might have more low quality reads, or another sample might have a slightly higher concentration on the flow cell.

Gene	Sample #1 635 reads	Sample #2 1,270 reads
A1BG	30	60
A1BG-AS1	24	48
A1CF	0	0
A2M	563	1126
A2M-AS1	5	10
A2ML1	13	26

Data normalization

- Each sample will have different number of reads assigned to it due to the fact that one sample might have more low quality reads, or another sample might have a slightly higher concentration on the flow cell.

Gene	Sample #1 635 reads	Sample #2 1,270 reads
A1BG	30	60
A1BG-AS1	24	48
A1CF	0	0
A2M	563	1126
A2M-AS1	5	10
A2ML1	13	26

Types of normalization

- **Fragments Per Kilobase Million (FPKM)**

FPKM is made for paired-end RNA-seq. With paired-end RNA-seq, two reads can correspond to a single fragment so it doesn't count the fragment twice. Here's how you calculate FPKM:

1. Count up the total fragment in a sample and divide that number by 1,000,000 – this is the "per million" scaling factor.
2. Divide the fragment counts by the "per million" scaling factor. This normalizes for sequencing depth, giving you fragments per million (FPM)
3. Divide the FPM values by the length of the gene, in kilobases. This gives you FPKM.

By Dr. Gang Wu: https://wiki.stjude.org/display/CAB/RNAseq+differential+gene+expression+analysis

Types of normalization

- **Fragments Per Kilobase Million (FPKM)**

FPKM is made for paired-end RNA-seq. With paired-end RNA-seq, two reads can correspond to a single fragment so it doesn't count the fragment twice. Here's how you calculate FPKM:

1. Count up the total fragment in a sample and divide that number by 1,000,000 – this is the "per million" scaling factor.
2. Divide the fragment counts by the "per million" scaling factor. This normalizes for sequencing depth, giving you fragments per million (FPM)
3. Divide the FPM values by the length of the gene, in kilobases. This gives you FPKM.

- **Count per million (CPM)**

CPM is calculated the same way as FPKM except it is not normalized by gene length, i.e. it skips the step 3.

By Dr. Gang Wu: https://wiki.stjude.org/display/CAB/RNAseq+differential+gene+expression+analysis

Types of normalization

- **Fragments Per Kilobase Million (FPKM)**

FPKM is made for paired-end RNA-seq. With paired-end RNA-seq, two reads can correspond to a single fragment so it doesn't count the fragment twice. Here's how you calculate FPKM:

1. Count up the total fragment in a sample and divide that number by 1,000,000 – this is the "per million" scaling factor.
2. Divide the fragment counts by the "per million" scaling factor. This normalizes for sequencing depth, giving you fragments per million (FPM)
3. Divide the FPM values by the length of the gene, in kilobases. This gives you FPKM.

- **Count per million (CPM)**

CPM is calculated the same way as FPKM except it is not normalized by gene length, i.e. it skips the step 3.

- **Transcripts Per Kilobase Million (TPM)**

TPM is very similar to FPKM. The only difference is the order of operations. Here's how you calculate TPM:

1. Divide the fragment counts by the length of each gene in kilobases. This gives you fragments per kilobase (FPK).
2. Count up all the FPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.
3. Divide the FPK values by the "per million" scaling factor. This gives you TPM.

** When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of fragment that mapped to a gene in each sample. In contrast, FPKM, the sum of the normalized fragments in each sample may be different, and this makes it harder to compare samples directly. **

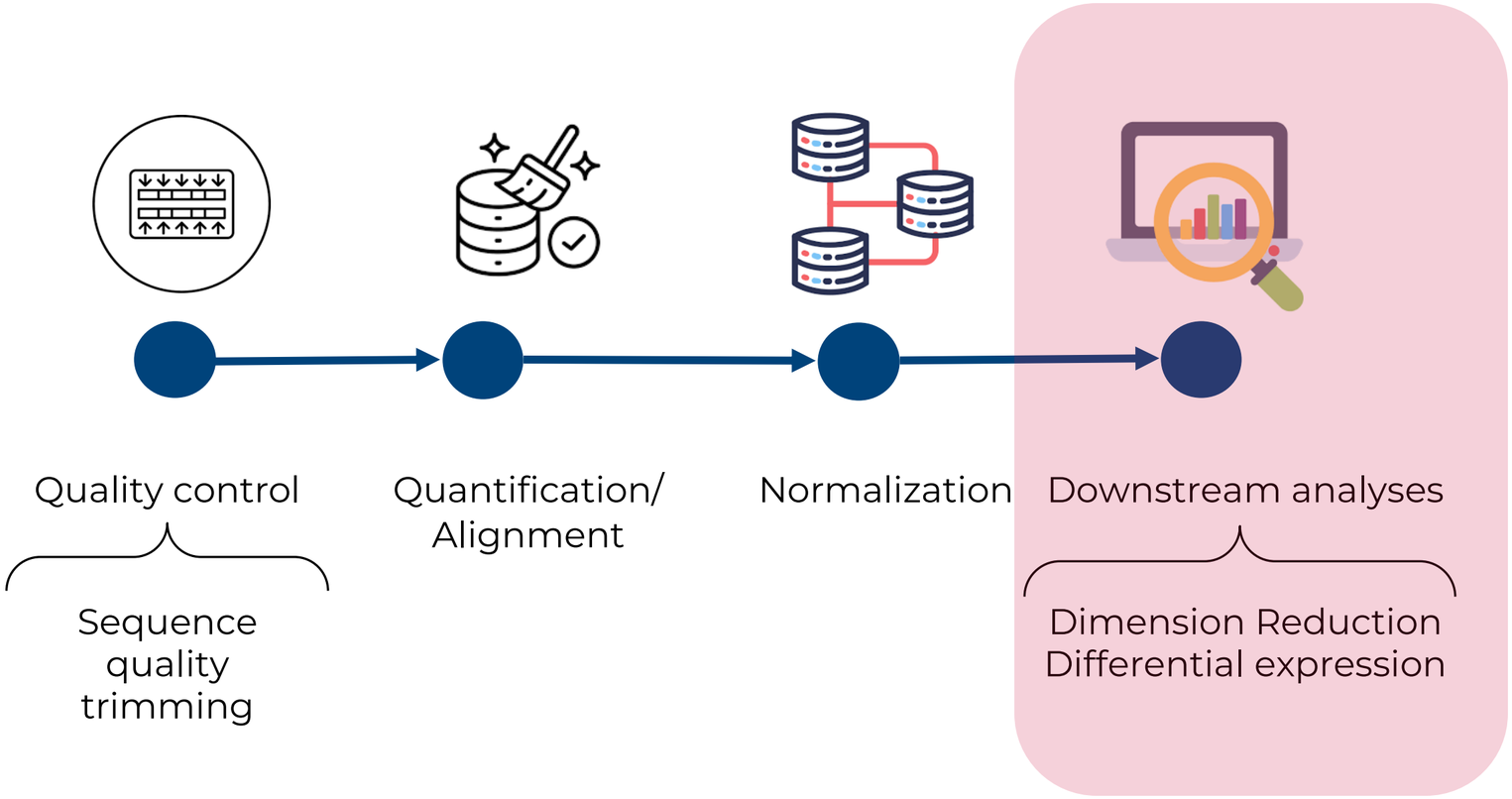
By Dr. Gang Wu: https://wiki.stjude.org/display/CAB/RNAseq+differential+gene+expression+analysis

Differential expression analysis (Limma voom)

Read counts per gene (raw count data) are analyzed using R software using the following methods:

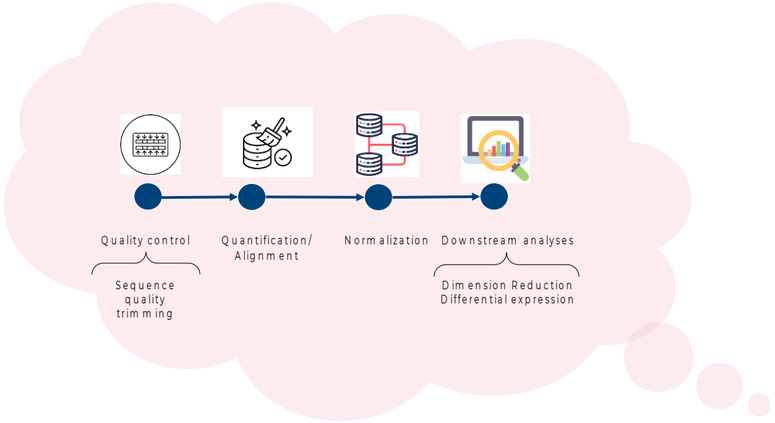
1. Only confidently annotated (level 1 (verified) and 2 (manually annotated) gene annotation)
2. Protein-coding genes are considered in the standard differential expression analysis.
3. Low count genes were removed from analysis using a CPM cutoff corresponding to a count of 10 reads.
4. Normalization factors were generated using the TMM method.
5. Counts were normalized using voom.
6. Voom normalized counts were analyzed using the lmFit and eBayes functions of the limma software package.
7. MA plots were generated from voom normalized data and volcano plots generated from voom normalized data plus p-values values from limma.

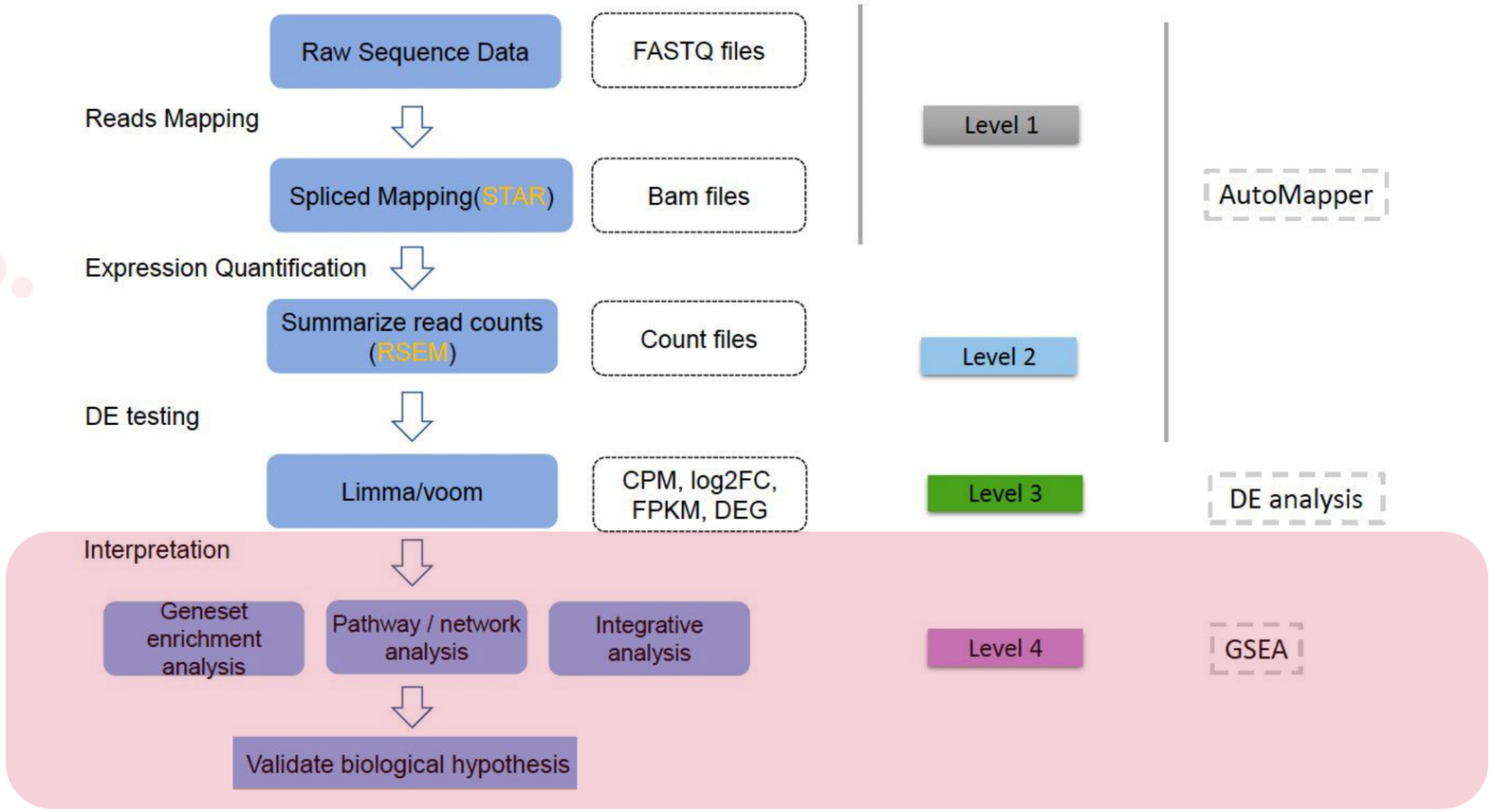
RNA-seq computational workflow



Source (edited): `Understanding Gene Expression Data` course by https://www.itcrtraining.org/home

Downstream analysis workflow





- R shiny app/RNAseqV2.3X
- Manual on how to use the app

Let us help you!

Providing advanced bioinformatic services for investigators to leverage omics data

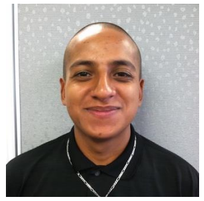
- Project analysis
- Benchmarking

- Training

- CAB liaison
- Consultation

- Innovation

GitHub



Cody Alexander Ramirez, PhD
Senior Bioinformatics Research Scientist
Core Director
Boston, Massachusetts

Antonia Chroni, PhD
Senior Bioinformatics Research Scientist
New York, New York

Sharon Freshour, PhD
Bioinformatics Research Scientist
St. Louis, Missouri

Asha Jacob Jannu, PhD
Bioinformatics Research Scientist
Indianapolis, Indiana